Image usefulness of compressed surveillance footage with different scene contents

Tsifouti, A.

IMAGE USEFULNESS OF COMPRESSED SURVEILLANCE FOOTAGE WITH DIFFERENT SCENE CONTENTS

ANASTASIA TSIFOUTI

A thesis submitted in partial fulfilment of the
requirements of the University of Westminster
for the degree of Doctor of Philosophy

This research programme was carried out within the
Computational Vision and Imaging Technology (CVIT)
research group at the University of Westminster

May 2016

Abstract

The police use both subjective (i.e. police staff) and automated (e.g. face recognition systems) methods for the completion of visual tasks (e.g person identification). Image quality for police tasks has been defined as the image usefulness, or image suitability of the visual material to satisfy a visual task. It is not necessarily affected by any artefact that may affect the visual image quality (i.e. decrease fidelity), as long as these artefacts do not affect the relevant useful information for the task. The capture of useful information will be affected by the unconstrained conditions commonly encountered by CCTV systems such as variations in illumination and high compression levels. The main aim of this thesis is to investigate aspects of image quality and video compression that may affect the completion of police visual tasks/applications with respect to CCTV imagery. This is accomplished by investigating 3 specific police areas/tasks utilising: 1) the human visual system (HVS) for a face recognition task, 2) automated face recognition systems, and 3) automated human detection systems.

These systems (HVS and automated) were assessed with defined scene content properties, and video compression, i.e. H.264/MPEG-4 AVC. The performance of imaging systems/processes (e.g. subjective investigations, performance of compression algorithms) are affected by scene content properties. No other investigation

has been identified that takes into consideration scene content properties to the same extend. Results have shown that the HVS is more sensitive to compression effects in comparison to the automated systems. In automated face recognition systems, 'mixed lightness' scenes were the most affected and 'low lightness' scenes were the least affected by compression. In contrast the HVS for the face recognition task, 'low lightness' scenes were the most affected and 'medium lightness' scenes the least affected. For the automated human detection systems, 'close distance' and 'run approach' are some of the most commonly affected scenes. Findings have the potential to broaden the methods used for testing imaging systems for security applications.

Contents

List of Figures

List of Tables

Acknowledgements

Successes never happen individually and this thesis was made possible due to the support received from numerous people. The greatest support has been received from my parents, apart from the emotional support, they have also provided significant financial support. I would like to thank the Home Office Centre for Applied Science and Technology (CAST) for sponsoring this work financially for the first 2 years and for allowing some time off as a civil servant to work on this research. I would like to thank work colleagues Mr Stuart Rankin, Dr Steve Bleay and Dr Andrew Barns for contributions in the experiments in Chapters 4 and 6. Special thanks to my colleague Mr Graham Doré for creating software that has been utilised in the experiment in Chapter 6. Additional thanks to Mr Fell David from Transport for London (TfL) for assisting in the execution of the methodology in Chapter 4.

I am the most grateful to my primary research supervisor Dr Sophie Triantaphillidou for guiding me through my PhD journey. Special thanks to my other 2 supervisor Dr Efthimia Bilissi and Dr Alexandra Psarrou for providing valuable scientific, technological and methodological input. I would like to express my sincere thanks to Dr Chaker Larabi from the University of Poitiers in France for a significant contribution to the produced publications. Also, exceptional thanks to my work line

manager at CAST Mr Anthony Clark for long discussions on statistical analysis matters and support received for many years. Because of the aforementioned people I have gained valuable academic skills relating to publication procedures and scientific reasoning.

This would have been unattainable without the support received form family and friends. Special thanks to Dr Jae Young Park, Colleen Addicott, Dr Gaurav Gupta and Daniel Broncano for providing support. My sister Sofia Tsifouti, brother in law Theo Kyfonidis, my niece Kiki Kyfonidis, and 2 nephews Damian and Ioannis-Chrisostomos Kyfonidis for adding fun in my life.

Dedicated to my parents

Kyriaki and Chrisostomos

Affirmation

This thesis was created at the Computer Vision and Imaging Technology (CVIT) research group of Westminster University and at the Centre for Applied Science and Technology (CAST) of Home Office Science UK.

I declare that the work submitted is my own. Any other sources of information or thoughts and visual material that have been used, directly or indirectly, have been referenced appropriately.

Anastasia Tsifouti

CHAPTER 1

Introduction

This research investigates aspects of image quality and video compression that may affect the completion of police tasks from Closed-Circuit Television (CCTV) imagery. CCTV imagery is used in UK courts as documentary evidence [1] and it has been found to have an effective impact on conviction of crimes [2]. When a crime occurs, police officers gather evidence (e.g. from the crime scene) and carry out recognition tasks to prove identities. The court makes identification decisions (i.e. establishes formally identities) from the available evidence. Often a combination of evidence (e.g. fingerprints, DNA, CCTV) are presented to the court for identification purposes [3]. Depending on the seriousness of a case (e.g. murder, rape, terrorist attack), a visually poor reproduced imagery (or any other type of poor evidence) might still be used as evidence or clue in an investigation. Often a 'poor evidence' might be used for elimination purposes from possible suspects. In exceptional cases a single piece of strong evidence might be adequate for the court to make an identification decision (e.g. a distinctive feature such as a person's gait, face or clothing) [4].

Figure 1.1 provides a CCTV example of a shop robbery [5]. The robbers in the scene have their facial information covered but clothing information is visible and quite distinctive; they are not wearing clothing of uniform colours. For example, one of the robbers is wearing a hoody top that includes American flags and the word California. The second robber is wearing black and white clothing and when the entire footage is viewed (i.e. not just this single image) then further information on his clothing can be obtained (e.g. words and patterns).

Figure 1.1: An example CCTV imagery of distinctive clothing. From Manchester Evening News website [5].

Information that the police can gather from the footage in Figure 1.1 relates to the type of weapon in the attack (i.e. in this case a machete), clothing (e.g. words, patterns, type and colour), the way robbers move or even identification of their gait (e.g. abnormal movements would be considered distinctive for example a limp), and calculation of their heights (i.e. by utilising photogrammetric techniques [6,7]). Additionally, one can observe from the left hand side robber in Figure 1.1 that the words on the hoody are not very clear due to the angle of the hoody to the camera plane. Yet, one can distinguish/recognise the word California even though the exact letters are not all legible. This is similar to how humans process known faces from poor quality CCTV video footage; the brain has the ability to put together memorised information combined with perceived information (i.e. the filling-in phenomenon [8]). For example, research has proven that humans have an excellent ability to recognise known faces (i.e. derived from memory of faces that were learnt minutes, hours, or years ago) even from high levels of degraded CCTV footage but

performance is reduced dramatically when the faces are unknown [9–13].

CCTV footage is used by law enforcements authorities for the completion of recognition tasks from visual information. These tasks relate to: 1) recognition of a person from facial, clothing, or gait information, 2) recognition of an action (e.g. who gave the first punch), and 3) recognition of an object (i.e. number plate, vehicle type) [14–16]. The aforementioned visual recognition tasks could be completed by utilising either humans (e.g. police officers, special analysts) and/or automated systems (e.g. automated face recognition, human detection systems).

The term image quality is utilised in the same manner between automated and human visual systems for the completion of police recognition tasks [14, 16–18]. For example, according to UK passport requirements, a good quality image for a human face recognition task should be: correctly exposed, include no occlusion such as glasses/hat, consist of a uniform background and convey frontal facial angle information (see Figure 2.4) [19]. Still, if the face includes distinctive characteristics (e.g. piercings, tattoos, birth marks, shape/size of nose) than a hat or glasses might not affect the human recognition task. As a result, the term *image usefulness* (or *image utility*) instead of image quality is adopted in the literature for police tasks. *Image usefulness* is associated with image quality and relates to the suitability of the imagery to satisfy a task [18]. The term image quality is often considered a general term due to the multiple applications and the broad nature and disciples relating to imaging systems [20]. For example, a portrait would be judged differently in the arts context (i.e. in terms of aesthetics) compared with a police application (i.e. visibility of appropriate information for a face recognition task).

High compression levels are favoured in the CCTV industry, since they allow more hours of recording and lower the cost of a storage system (and transmission). However, they compromise the image usefulness of the recorded imagery. Video compression in the security industry employs proprietary formats based on industry standard compression algorithms. The H.264/MPEG-4 AVC algorithm has been identified to represent a popular, current and future trend in the CCTV indus-

try [21–23]. H.264/MPEG-4 AVC is a hybrid video encoder, exploiting both spatial and temporal redundancy utilising an approximation of the 4×4 Discrete Cosine Transfer (DCT). The H.264/MPEG-4 AVC compression algorithm is investigated in this thesis. This compression algorithm produces blocking artefacts that become more visible at high compression levels [24]. Compression artefacts will not necessarily affect the image usefulness of CCTV imagery, as long as these artefacts do not affect the relevant visual information to the recognition task.

Figure 1.2: Reduction of usefulness of information from the reference scene due to wavelet compression.

Figure 1.2 illustrates a CCTV example of a car park and how image usefulness can be judged subjectively from the entailed information. The top left image represents the reference 'uncompressed' scene and the other 3 scenes represent compressed versions of the reference ranging from low (i.e. light) compression to medium and high. The reference scene illustrates the maximum available information that can be captured by the system under those specific conditions (e.g. camera to subject distance, illumination conditions). The compressed scenes are degraded version of the reference. In the low compressed scene almost all of the useful information has been maintained from the reference and in the highly compressed scene there is information loss on clothing patterns, facial detail, colour and general definition

of shapes. If observers are asked to judge the image usefulness of the compressed scenes then they will be confused, as the image usefulness is dependent on the police task. For example, if the police task is to count the number of people in Figure 1.2 then even the highly compressed version is good enough for the completion of that task. On the other hand, if the police task is to recognise someone from clothing information then the highly compressed version is not good enough for the completion of that task. For this reason, this research includes 3 specific police tasks that are investigated as case studies in connection with CCTV imagery and compression. These specific police tasks are linked with: a) human face recognition, b) automated face recognition, and c) automated human detection as part of a video analytics (VA) system. As it has been described in the previous paragraphs the term image usefulness is utilised in the same way for both human and automated systems for police tasks. The inclusion of 2 investigations (human and automated) for the same task (i.e. face recognition) will allow the identification of correlations between them.

Furthermore, the aforementioned police tasks were chosen in order to cover a wide spectrum of police applications utilising both human and automated visual systems. Faces are a non-intrusive biometric, the most exposed (in comparison to fingerprints), used across many security applications (e.g. police, borders), and the most available (e.g. on social websites, police mugshots, passports, ID cards, street CCTV cameras). There are many surveillance CCTV applications where the capture of facial information is possible, such as in trains, buses, underground, transport stations and open street. The automated human detection task belongs to the video analytics (VA) systems category. VA are autonomous systems (i.e. with little or no human interaction) [25] with the aim of replacing the monotonous task of human visual examination of video data. VA systems might become the future of policing but currently very few scenarios are capable of autonomous analysis. One of the capable scenarios is the sterile zone (SZ), which is investigated in this thesis [26, 27]. The SZ scenario (see Figure 2.9) consists of a fence and an area with grass (not to be trespassed); the VA system needs to alarm when there is an

intruder/human entering the scene. The task is automated human detection.

Image usefulness can be obtained from measuring performance of systems such as results denoting correct (or missed or false) detection/recognition utilising either human or automated visual systems. For example, recognition of an unknown facial image among a dataset of known facial images is considered correct recognition. Further, when assessing human visual systems (e.g. police officers), experience derived from completing recognition tasks can also be utilised to further understand appropriateness of image usefulness. For example, police officers can be asked if the highly compressed scene in Figure 1.2 is good enough for counting the number of people; if the answer is yes then the compression amount applied is acceptable for that task.

To date there has not been much direct research carried out on the subject of image usefulness of CCTV imagery. Resources for human face recognition tasks come mainly from psychology for face recognition and the exhaustive work by Klima [28] and his co-authors on different compression techniques and their impact on CCTV footage. Klima and co-authors [15,28–30] tested many different compression techniques using subjective testing. They concluded that the perceived quality was not dependent just on compression rate, but on the initial information content of the scenes and its purpose [29]. Their work is related mainly to a few close up faces and number plates. For this reason, in the present investigation, a more extensive set of scenes with different attributes and properties is included.

The findings in automated face recognition using still imagery with JPEG [31] and JPEG2000 [32] standards agree that compression does not adversely affect automated face recognition performance [33–37]. In a study by the Face Recognition Vendor Test (FRVT) with JPEG compression [33], the findings have shown no deteriorated performance with images compressed to 0.2bpp (bits per pixel) [33] and performance deteriorated under 0.2bpp. The same study has even reported an increase in performance of face recognition systems with compressed images to 0.8bpp and 0.4bpp (values above 0.2bpp) [33]. These results are consistent with another

investigation conducted by Delac and co-authors [35,37]. In most cases, face recognition systems are evaluated based on their performance of correct recognition from large datasets [33,38–40] and individual properties of each facial image (e.g. over or under exposed faces) are not considered. For example, images obtained from a web-camera will just be labelled as being of poor quality [41]. Furthermore, Aggarwal et al. [42] have identified that face recognition systems performed differently on different datasets due to the dissimilar scene properties of the facial imagery under each dataset. For example, results of performance comparison between PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) face recognition holistic techniques differ significantly among different databases and no conclusion can be made on the best performing one [43]. Face recognition holistic techniques utilise the whole face region rather than individual features (e.g. eyes, mouth). In this thesis, face properties are characterised (e.g. how far away is the face from the camera plane) in order to identify the scene properties contributing to the decrease in performance of automated face recognition systems.

Additionally, little research has been accomplished in identifying relationships between human evaluators and face recognition algorithms in relation to face image usefulness. An investigation by Adler and Dembinsky [44] has found that derived biometric image quality scores do not correlate between human perception and automated algorithms. This thesis will investigate correlations between human and automated visual systems for the face recognition task.

Poppe *et. al.* [45] have pointed out the lack of research in the area of video analytics systems with compressed footage. The same authors have tested a background subtraction technique based on the Gaussian Mixture Models (GMM) with the H.264/MPEG-4 AVC video coding standard. The performance of GMM was not dependent only on the amount of compression but also on scene content. In another investigation [46] with the SZ scenario (or intruder detection) and H.264/MPEG-4 AVC video coding standard, the results have shown the performance of the analytics system to be affected at 220kbps (kilobits per second), either by not detecting an

attack, or producing a slower alarm response time. That work investigated 11 attacks with only 1 VA system. In this thesis, a much greater amount of footage (or attacks) and number of VA systems are investigated.

As an extra factor, results from subjective investigations and assessments of compression performance are often shown to vary with image content [29,30,47,48]. For example, in video, scenes with different spatial (varied regions)-temporal (varied motions) structural properties will require different bit-budgets leading to different levels of compression.

In conclusion, any visual process (subjective investigations, compression algorithms, automated recognition/detection systems, human face recognition) is dependent on scene content. Scene dependency can be overcome with the use of scene characterisation and classification methods [47]. Scene contents can be characterised and later classified to groups of certain scene characteristics/properties. Both objective (use of relevant algorithms) and subjective methods (visual or empirical inspection) have been employed in this thesis for characterisation purposes. For example, scene lightness of facial imagery was deduced objectively from measuring skin lightness using the L^* value from the CIELAB colour space. Skin lightness denotes if the scene is under, over, mixed or correctly exposed. In opposition, angle of face to the camera was deduced subjectively by visual inspection.

1.1 Aims and objectives

The main aim of this research is to investigate aspects of image quality and video compression that might affect the completion of police visual tasks with respect to CCTV imagery. The following objectives are going to be achieved.

- to contribute to increasing understanding in the area of image usefulness for police visual tasks;
- to identify appropriate methodologies for testing/assessing police visual systems with CCTV imagery and video compression;

- to develop a dataset representative of the challenges encountered by real-world CCTV applications for face recognition investigations (both human and automated);
- to identify the extent to which compression (H.264H.264/MPEG-4 AVC) affects the completion of police tasks for the 3 aforementioned applications under investigation;
- to identify differences/similarities between the industry-standard compression algorithm H264/MPEG-4 AVC and proprietary format versions employed by CCTV systems;
- to identify the image content characteristics/properties that will affect the completion of the 3 aforementioned police applications in combination with and/or without compression;
- to identify any relationships between human and automated face recognition systems in relation to scene content properties and compression.

These objectives were achieved by assessing 2 types of visual systems (human and automated) for 3 specific police tasks (human face recognition, automated face recognition, automated human detection) with characterised CCTV imagery and video compression (H.264/MPEG-4 AVC). Knowing exactly with what content characteristics a system (e.g. automated systems, compression algorithms) fails can contribute to the further improvement of such a system. For example, to allow compression up to an acceptable level for the human face recognition task.

For the face recognition task a new dataset was developed in order to utilise test material that comes from a challenging CCTV application. This application was the London bus; the window features on buses create challenges in terms of illumination. For example, on a sunny day when the bus is in motion, windows allow illumination from different directions, causing areas of over and under exposure. On the other hand, during night conditions, the bus illumination is the principal source; it provides relatively uniform illumination. Also, around 10 cameras were installed

on a London bus allowing the capture of content with varied properties. For example, the camera installed for viewing the staircase (i.e. at the top deck) is ideal for capturing tilted angle faces. Whereas, the camera installed on the back window (i.e. for both decks) is ideal for capturing frontal angle faces. Figure 4.2 presents an example of the captured camera views. Further, the already available face datasets are created in semi or fully controlled environments [33, 38–40], which might not be as pragmatic as the dataset created from the London bus application.

In case of the automated human detection task, an already available dataset was utilised. The name of the utilised dataset is the sterile zone (SZ) scenario, which is part of the Imagery Library for Intelligent Detection Systems (iLIDS) datasets [49, 50]. iLIDS is a UK government initiative that provides to the manufacturers of VA systems, datasets with a wide variety of scenarios in relation to police tasks (e.g. detection of abandoned baggage in London tube, detection of prohibited parking of vehicles). The manufacturers develop systems based on these provided scenarios. The iLIDS team benchmarks the performance of VA systems and provides to the manufacturers a UK government classification standard. It was considered important for the human detection task to include the SZ scenario dataset as the systems under assessment are designed for this specific scenario.

The following section presents some further information on the content of the thesis. The publications arising from this work can be found in Section 1.3. Section 1.4 discuses the original and significant contributions to knowledge of this research.

1.2 Content of the thesis

Chapters 2 and 3 include all the background information necessary to enable an insight to the factors that have to be taken into consideration before developing appropriate evaluation performance methodologies for the visual police systems under investigation. Chapter 2 provides information in relation to the applications of CCTV imagery and police tasks. Later, the chapter focuses on the 3 investigated

applications: human face recognition, automated face recognition and automated human detection (i.e. as part of VA systems). Further, the challenges affecting the task performance of the police visual systems are described. Chapter 3 provides information on the subjects of video compression and image quality with respect to CCTV imagery. This information includes details on: video compression for both 'standard' and CCTV industry, definitions of image quality and image quality attributes, methods in psychophysical investigations and the factors that affect such procedures, and scene content characterisation and classification.

Chapter 4 identifies acceptable compression limits (relating to image usefulness) for human face recognition, using psychophysical investigations, an industry implementation of the standard H.264/MPEG-4 AVC, the CCTV recording systems on London buses and a variety of scene content properties. The London bus application is utilised as a case study for setting up a methodology and implementing suitable data analysis for face recognition from recorded footage, which has been degraded by compression. The footage has been characterised and classified to predefined scene groups relating to skin lightness (i.e. this determines if the scene/face is under, over, mixed or correctly exposed), camera to subject distance (i.e. a close distance scene will reveal more facial information from a further distance scene), facial angle to the camera plane (e.g. frontal and tilted), and level of busyness (i.e. based on spatial and temporal information). Psychophysical investigations are conducted in order to transform subjective judgements to quantitative results. In these investigations the compressed version(s) are judged against its 'uncompressed' reference (i.e. a similar procedure to the example in Figure 1.2 is followed). The analysis of the results is based on the individual scene properties (e.g. close, far distance to the camera plane), type of observers (i.e. police officers, surveillance officers and bus analysts) and type of compression algorithm (i.e. industry-standard or CCTV proprietary format). Additionally, a section on the development of the representative video dataset from the London bus application is provided.

Chapter 5 provides a comparative performance evaluation between human and auto-

mated face recognition systems, using CCTV imagery, different compression levels and scene properties. Results and test material obtained from the human investigation are also utilised here. The aim of this investigation is to identify relationships between human and automated face recognition systems with respect to compression. Further, to identify and compare the most influential scene properties on the performance of each face recognition system. The investigation includes 3 basic automated face recognition (AFR) systems [51]: a) Principal Component Analysis (PCA), b) Linear Discriminant Analysis (LDA), and c) Kernel Fisher Analysis (KFA). The results are analysed using a distance measure between a degraded/compressed image from its reference 'uncompressed' version, which complies with the methodology utilised for the human investigation in Chapter 4.

Chapter 6 investigates the effects of scene content properties, frame rate and video compression on the performance of automated human detection systems with the SZ scenario. In this thesis 4 detection systems are tested with compressed (at 5 and 25 frames per second) and 'uncompressed' (only at 25 frames per second) footage of the SZ scenario. The scene properties were extracted from the characterisation of the content of 110 attacks (scenes). The characterisation included both objective and subjective techniques relating to scene contrast (contrast between main subject and background), camera to subject distance, subject description (e.g. 1 person, 2 people), subject approach (e.g. run, walk), and subject orientation (e.g. perpendicular, diagonal). Additional footage, including only distractions (e.g. foxes, birds, and weather conditions such as snow and rain) and no attacks to be detected is also investigated. The analysis of the results is based on identifying correct, missed or false detection for each individual grouped scene property under investigation. For example, the analysis can identify the proportion of correct detection for the human run scene property.

Chapter 7 provides an in-depth discussion on the obtained results from the 3 investigations. Lastly, in Chapter 8, conclusions are drawn with recommendations for further work.

1.3 Produced publications

Journal Articles:

A. Tsifouti, S. Triantaphillidou, M.-C. Larabi, G. Dor, E. Bilissi and A. Psarrou. **A case study in identifying acceptable bitrates for human face recognition tasks**, Elsevier, Signal Processing: Image Communication, 36(0), 14-28p, (2015).

Proceedings Articles:

A. Tsifouti, S. Triantaphillidou, M.-C. Larabi, G. Dor, E. Bilissi and A. Psarrou. **The effects of scene content parameters, compression, and frame rate on the performance of analytics systems**, Proc. SPIE 9396, Image Quality and System Performance XII, 93960X (January 8, 2015).

A. Tsifouti, S. Triantaphillidou, M.-C. Larabi, E. Bilissi and A. Psarrou. **Comparative performance between human and automated face recognition systems, using CCTV imagery, different compression levels and scene parameters**, Proc. SPIE 9396, Image Quality and System Performance XII, 93960M (January 8, 2015).

A. Tsifouti, S. Triantaphillidou, E. Bilissi and M.-C. Larabi. **Acceptable bitrates for human face identification from CCTV imagery**, Proc. SPIE 8653, Image Quality and System Performance X, 865305 (February 4, 2013), **(Obtained Best Student Paper award)**.

A. Tsifouti, M. M. Nasralla, M. Razaak, J. Cope, J. M. Orwell, M. G. Martini and K. Sage **A methodology to evaluate the effect of video compression on the performance of analytics systems**, Proc. SPIE 8546, Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII, 85460S (October 30, 2012).

1.4 Original contributions to knowledge

This research has contributed with original knowledge by:

1. Investigating for the first time image usefulness (an attribute that has its origins in Yendrikhovskijs Fidelity Usefulness Naturalness framework and relates to the suitability of the imagery to satisfy a task [18]) in the context of imagery relating to security/police applications. In this thesis, image usefulness has been defined for three specific police applications relating to i) human face recognition, ii) automated face recognition and iii) automated human detection.

2. Identifying and quantifying original scene properties (e.g. scene illumination, spatio-temporal business) or/and facial properties (e.g. tilting facial angle, camera to subject distance) that contribute to the successful and/or unsuccessful completion of all three police tasks mentioned in 1. These new findings have the potential to contribute to the development, and particularly to the parametrisation, of image quality metrics used in police tasks.

3. Developing an original image dataset, the CASTBUS 2012 dataset, using the understanding that scene content and facial properties affect human face recognition results. The CASTBUS 2012 dataset includes footage with varied scene content properties in terms of: camera to subject distance, spatio-temporal busyness, illumination conditions and facial angles to camera plane. This dataset is available to those researching in relevant areas.

4. Providing novel experimental paradigms for testing imaging systems relating to security/police. All methodologies included in this thesis are carefully thought/applied, by combining resources from different academic scientific disciples (relating to image processing and compression, visual psychophysics, image quality, use of police imagery, face recognition studies and automated systems/algorithms) to identify research solutions. This is a significant contribution that will allow improvements in testing police systems (humans and

automated).

5. Providing useful results to relevant communities. For example, Transport for London (TfL) has implemented the compression recommendations derived from the human face recognition investigation in a London bus. This will result to having more suitable recording and compression conditions (i.e. acceptable compression levels) for face recognition tasks undertaken by specialists.

Overall, little research has been accomplished in the area of image quality/usefulness for police tasks. This thesis widens the understanding of particular topics/issues by providing an imaging scientists perspective. For example, investigations in human face recognition have been predominately accomplished by psychologies, or neuroscientists who do not necessarily account for the effects of imaging properties/attributes in the task. The same applies to computer vision scientists who produce automated recondition systems. For instance, the majority of footage/still image datasets that have been created over the years from the computer vision community do not take scene content properties into a consideration and often standard datasets are created without controlled variation of such information.

CHAPTER 2

Applications of CCTV imagery

This chapter provides information in regards to closed-circuits television (CCTV) imagery for law enforcement purposes. Additionally, police applications relating to CCTV imagery are described. Later, the chapter focuses on providing some background information on 3 specific police applications, which are the ones investigated in the experimental part of the thesis. The applications relate to human face recognition, automated face recognition and automated human detection (i.e. as part of video analytics systems). Also, the challenges influencing task performance of human operatives (i.e. police officers) and automated systems (i.e face recognition, video analytics) with CCTV imagery are described. For example, CCTV systems often operate under totally uncontrollable, or semi-controllable illumination conditions (e.g. open street CCTV cameras, bus CCTV systems) that might affect the capture of useful information (e.g. face). The usability of the imagery is further compromised by compression, implemented to satisfy limited storage capacity of CCTV recorded systems, or transmission bandwidths. Low bitrates are favoured in the CCTV industry for lowering data costs.

2.1 CCTV imagery

The breakthrough for the use of CCTV systems happened when the solid-state CCTV cameras, such as charge-coupled device (CCD) cameras, were introduced in the early 1990s [52]. These solid-state cameras required minimal maintenance. There are 2 studies that provide estimations on the number of CCTV cameras in the United Kingdom (UK). The first study was conducted in 2003 with an estimated number of 4.2 million CCTV cameras [53]. A more resent study in 2011 gave an estimation of 2 million [54]. These are just estimations and do not represent the actual number of the CCTV cameras in the UK. Nevertheless, one can conclude that there is a vast amount of video CCTV data that can be used to prevent and solve crime. A couple of main factors have contributed to the widespread use of such systems: one of them is the increase of crime (including terrorist attacks) [55] and the second one is the current availability of advanced and low-cost systems [52].

A Home Office study has found CCTV cameras ineffective in terms of reducing crime but having an effective impact on the handling of individual incidence and high profile cases [2]. The effectiveness of the CCTV systems will be compromised when the wrong cameras are fitted (i.e. if the location is wrong or if the cameras are not working), when the operators are not trained to handle such systems, and when the produced imagery is of low quality [2].

CCTV footage is used by the police for the completion of 3 main tasks: recognition of i) a person (i.e. from facial information, clothing, gait), ii) an action (e.g. who gave the first punch), and iii) an object (i.e. number plate, vehicle type) [14–16]. The terms *recognition* and *identification* are often used interchangeably in academic papers. The Oxford dictionary [56] defines recognition as the "identification of a thing or person from previous encounters or knowledge" and identification as the "means of providing a person's identity, especially in the form of an official paper". The Cambridge dictionary [57] defines recognition as the "fact of knowing someone or something because you have seen or heard him or her or experienced it before"

and identification as "the act of recognising and naming someone or something". According to these meanings, identification is formalisation of recognition. For example, you recognise a person and latter you define/establish that persons' identity. Putter [58] has pointed out the incorrect usage of the term identification for criminalistics by stating "In the field of criminal investigation, the general use of the word 'identification' differs markedly from the classical philosophical concept, since 'identity' itself is differently defined. Identification is the placing of an object in a class or group. This is the sense in which the word is used in all the natural sciences and to use it in any other sense in criminalistics is non-scientific. This scientific usage does fly in the face of popular practice, in which a criminal is 'identified' from his fingerprint. He is not identified, he is individualised...What is proved by his fingerprint is his individual identity, i.e. his individuality".

Police staff carry out recognition tasks and the court decides whether or not to convict based on the evidence provided. The court is responsible for making an identification or individualization decision. CCTV imagery is used in court as documentary evidence [1]. Other documentary evidence includes photographs, drawings, plans and maps. Often, CCTV evidence is used in combination with other evidence (e.g. witnesses, fingerprints, blood analysis and more). Evidence presented to the court is judged according to its weight in terms of how much it proves or disproves a case. For example, individual evidence might provide moderate support for conviction, but when viewed in combination with other evidence it might provide a strong support for a conviction [3].

In order for the CCTV footage to be viewed as strong evidence, the information provided in the footage must be exceptionally distinctive. For example, in a burglary case the perpetrator was identified by his distinctive way of walking (bowed legs and an unusual way of moving his left hand) from a poor quality CCTV footage [4]. Figure 2.1a shows an image from the CCTV footage on the night of a burglary where only the silhouette of the perpetrator is visible. Figure 2.1b shows a custody image of the suspect. A podiatrist derived information from the poor CCTV

footage and provided evidence in court of the strong similarity of gait between the perpetrator and the suspect (see Figure 2.1). The CCTV footage would not have been appropriate for face or clothing recognition tasks due to the lack of relevant visual information, however it was still able to provide 'exceptionally distinctive' evidence.

Figure 2.1: Distinctive gait in CCTV footage. An example of a) a CCTV image (left image) matched against b) the suspect's gait (right image). From Nixon *et al.* (2010) [4].

The police employ both subjective and automated tools for completing recognition tasks from CCTV imagery. Automated tools are automated systems, such as face recognition (FR), number plate recognition (NPR) and video analytics systems (e.g. detection of intruders) [59]. Subjective tools involve visual examinations of recorded/transmitted imagery, carried out by operatives, such as police staff and external specialists. Often, automated and subjective police tools are used in combination. For example, the input to a FR system is an unknown face, the system compares the unknown face with a database of known faces (i.e. the enrols) and returns back possible matches to the unknown face [60]. The operator then needs to verify and make a recognition decision on the returned matches.

In most cases, the information within the CCTV footage is compared and matched by visual assessment of similarities with either a real subject, or a real object, or another source of imagery (e.g. a mugshot). Figure 2.2, illustrates an example where an image from the CCTV footage is compared and matched with full-face

photographs of 4 individuals. It is obvious that the inadequate (long) subject to camera distance reveals information of the target's face, which may be insufficient for a face recognition task.

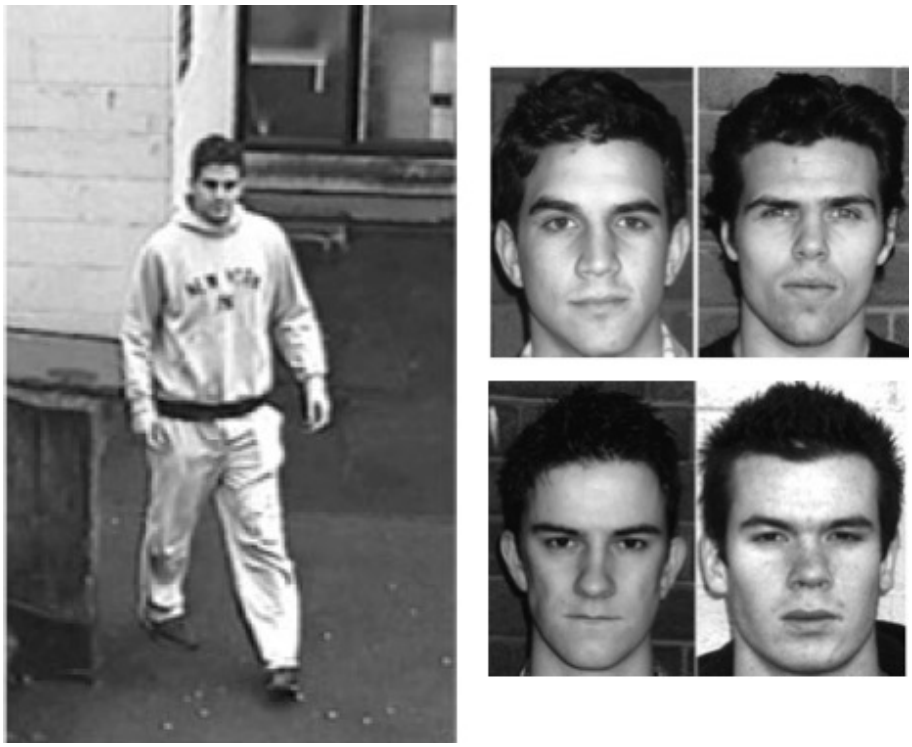


Figure 2.2: Face matching example. An image from the CCTV footage (on the left) where is compared and matched with full-face photographs of 4 individuals. From Davis *et al.* (2009) [61].

Furthermore, image matching techniques can also be used with scenes from CCTV imagery and its reconstructed versions. For example, in a court case a perpetrator's car in a CCTV scene was matched against a reconstructed scene with the suspect's car. The scene reconstruction was based on the information in the CCTV footage (i.e. positioning of the car and time of filming) and the use of the same CCTV system (i.e. same CCTV camera quality, system compression level and angle of the camera to the object). Imaging analysts used cues such as appearance of headlights and tax disc holder shape in order to make a decision if the perpetrator's and suspect's car are the same [3].

This thesis is concerned with police tasks relating to human face recognition, automated face recognition and video analytics systems (i.e. intruder detection in a sterile zone scenario). The following subsections provide some background informa-

tion on the 3 aforementioned police tasks.

2.1.1 Human face recognition

Research involving facial recognition tasks indicates that individuals have an excellent ability to recognise known faces (i.e. derived from memory of faces that were learnt minutes, hours, or years ago) even from high levels of degraded CCTV footage, but performance decreases dramatically when faces are unknown [9–13]. In image matching tasks the face is normally unknown and the comparison is based on 2 (or more) present stimuli, and not on memory and a present stimulus [13].

Factors such as illumination conditions, angle of the face to the camera plane, camera to subject distance and the physical size of printed images affect the accuracy of face matching tasks [11, 12, 62–66]. These factors affect the image quality of the reproduced imagery thus the capture of useful facial information. Figure 2.3 illustrates how facial information is affected by varying illumination conditions. Illumination poses an important problem in CCTV imagery [67], since CCTV systems often operate under totally uncontrollable, or semi-controllable illumination conditions (e.g. street CCTV cameras, bus CCTV systems).

Figure 2.3: Facial information under various illumination conditions. From Li *et al.* (2007) [68].

Image matching techniques are used for facial information comparisons, which are known as facial mapping. Facial mapping techniques fall into 3 categories: a) morphological (classification of features based on shape) [69], b) superimposition (overlaying of images) [70], and c) photoanthropometry (the use of facial landmarks as proportionality indices) [71]. Trained anatomists, anthropologists and facial mapping experts carry out such techniques. These techniques are used to identify sim-

ilarities and dissimilarities of facial information between 2 imaged faces in specific areas, such as the mouth, upper lip and chin [72]. The greater the similarities of the facial information between 2 images the greater the possibility of a match. Facial mapping experts also provide the factors that may affect the reliability of the method such as image quality factors (i.e. pixilation and illumination variations), the possibility of 2 people appearing indistinguishable and the lack of a facial feature database (e.g. statistical explanation) [72]. Facial mapping techniques have been judged to be subjective and results are often taken as non-scientific evidence, mainly due to the absence of a standard facial database and a lack of knowledge of the number of people with particular features [73]. Despite those arguments, facial mapping techniques have been used as evidence in UK courts since 1989 [74].

As already noted, CCTV footage is matched with other sources of imagery (e.g. police mugshots, passport photographs). Requirements for UK passport photographs and police mugshots provide specific instructions on image capture to optimise human face examination tasks [19, 75]. For example, UK passport photographs need to have a uniform background, the subject to be looking straight to the camera and the crown of the head to the chin to occupy a certain size in the picture [19]. No information is given on compression for passport photographs. Figure 2.4, shows an example of appropriate and inappropriate passport photographs.

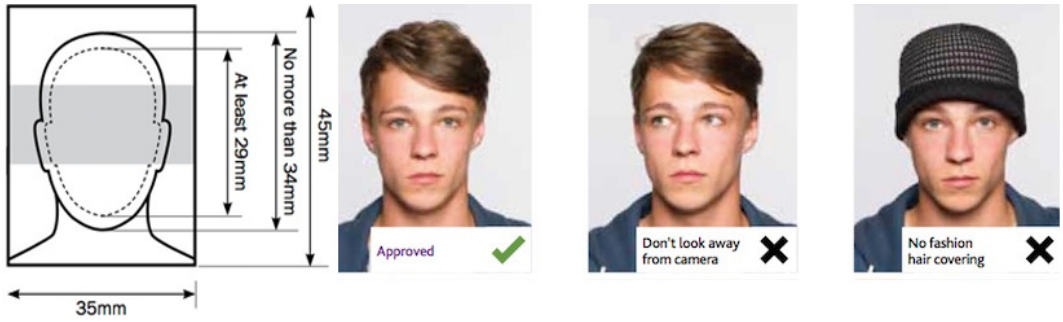


Figure 2.4: Example of UK passport photo requirements. From UK Government (2013) [19].

Police mugshot recommendations specify the compressed file size (in addition to head pose, illumination, uniform background, exposure and so on). They recommend the file size of an individual image to be between 24 and 44 kilobytes [75],

using JPEG or JPEG2000 compression. It is not known whether these recommendations were made based on empirical experience, or subjective investigations (see Section 3.4). Nonetheless, Figure 2.4 illustrates the human perception of a good quality facial image for a recognition task, which is: correctly exposed, no occlusion such as glasses/hat, uniform background, and of frontal angle.

The National Policing Improvement Agency (NPIA) [76] specifies that "JPEG and JPEG 2000 formats allow very good compression of photographic images, with minimal deterioration (artefacts) appearing in the image. This allows human examiners to see the mugshot image clearly, and automated face recognition systems to function effectively". In addition, NPIA [76] has defined the mugshot image as "an electronic colour image based representation of the portrait of a person with sufficient resolution for human examination as well as reliable computer facial identification. The image includes the full head with all hair in most cases, as well as neck and shoulders".

All the information in this section points to the priority of the police to capture useful image information, where deterioration due to image artefacts may be allowed, as long as the recognition tasks can be completed.

Section 3.3 identifies the term image quality to be strictly subjective as humans are the end users of imaging applications/systems/processes. There has not been much direct research carried out on the subject of image quality for CCTV systems. Resources come mainly from psychology for face recognition, event recognition, ergonomics on comfort of reviewing CCTV footage, military applications (most of the information is restricted), and an extensive work by Klima and his co-authors on different compression techniques and their impact on CCTV footage. Klima and co-authors [15, 28–30] tested many different compression techniques, using subjective testing indicating that H.264/AVC is a superior compression technique to MPEG-2 and DivX [30]. However, they have not used an extensive set of scenes, and they have not included different scene properties (e.g. different illuminations). Their work is related mainly to a few close up faces and number plates. Additionally,

they concluded that the perceived quality was not dependent just on compression rate, but on the initial information content of the scenes and its purpose [29]. For this reason, in this compression investigation a more extensive set of scenes with different attributes and properties is included and not just few scenes of close up faces.

2.1.2 Automated face recognition systems

Automated face recognition systems belong to the field of biometrics [77]. Biometrics, is the process of recognition and/or verification of an individual based on physiological (e.g. fingerprints, hand geometry) and/or behavioural (e.g. signature, voice) characteristics [78]. Behavioural characteristics are supposed to be learned over time and are subject to deliberate alternation, whereas physiological (i.e. physical) characteristics are more difficult to be manipulated/altered. For example, physiological characteristics are unique traits incorporated within each individual (i.e. they are individualised traits). There is a difference between verification and recognition when applied to automated systems. Verification (1 to 1 matching) is the combination of unique recognisers (or authentication modes) such as ID number and that person's biometrics. On the other hand, recognition (1 to many matches) utilises biometric measurements that are compared to a dataset of enrolled individuals (e.g. faces). In terms of government applications (e.g. to control immigration and prevent/solve crime) automated biometric systems are becoming of great significance.

This thesis deals with 2-dimensional (2D) face recognition in an unconstrained environment, which is where a common CCTV system will be operating. Unconstrained conditions create variations in facial images caused due to illumination (e.g. under or over exposed scenes), angle of face to the camera plane, distance of face to the camera plane, facial expressions and many more. Face properties are easier to use than other biometrics (e.g. finger, hand, voice) as they are non-intrusive and the most exposed [79]. In general, face recognition systems (automated and human)

work better for applications when training/enrolled and test/query facial images are captured under similar/controlled conditions (e.g. verification process of electronic passports at airports) [17, 80].

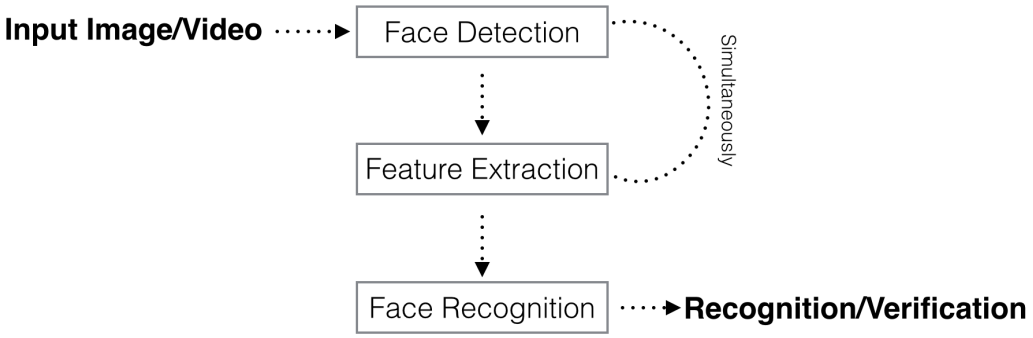


Figure 2.5: Generic processes of a face recognition system. Adopted from Zhao and Chellappa (2006) [81].

Figure 2.5 demonstrates the steps of a generic face recognition system. Generally, an interactive relationship exists between the steps. For example, facial features (eyes, noise, mouth, ears) might be used for both face recognition and face detection steps, whereas feature extraction and face detection can be operated at the same time [82, 83]. Face recognition has attracted a huge amount of interest from both academia and industry for the last 30 years. The result, is the development of many face recognition techniques [84–86]. These techniques accommodate the way humans process faces. For example, humans process faces using holistic (whole face region), featural (shape of individual features: eyes, nose and mouth), and configural (spacing among features: distance between chin and eye) information [87–90].

Similarly to the way humans process faces, algorithmic techniques of face recognition systems fall into 3 main categories:

- Holistic techniques. This approach creates face descriptors based on the whole face region rather than on shape/distance of individual features. The techniques can be classified into 2 further categories: statistical and artificial intelligence (AI). Statistical techniques employ dimensionality reduction methods. One of the most widely-used statistical technique is the eigenpicture based on the principal component analysis (PCA) [91, 92]. AI use artificial neural net-

works (ANN) and machine learning (ML) techniques. ML techniques learn a task from examples by executing automated actions based on binary and/or logical operations. ML techniques try to mimic human reasoning [93]. On the other hand, ANN techniques try to mimic the networks of neurons in the human brain. Neural networks consist of layers of interconnected and inter-dependent nodes, each node outputs a non-linear function from inputs (from other nodes or data inputs). [94].

- Feature-based techniques. This approach creates face descriptors based on shape and distance of individual features. Facial features (e.g. eye, mouth, noise and fiducial marks) are first extracted and later their geometric relationships are modelled. Pattern recognition techniques can be used to match faces using those geometric relationships. Multiple feature-based techniques exist. A method of localising points (corners of eyes) in facial images is presented by Belhumeur *et al.* [95]. Feature-based techniques invoke 2 main disadvantages: difficulties in feature detection and extraction, and the chosen feature set might not have discriminating power between faces [96, 97]. The advantage of these techniques is that, in principal they should be invariant of scale and pose. Holistic techniques are not robust with pose changes (prefer frontal faces where both eyes are visible) and they normally apply an alignment stage before the enrolment of the face image for recognition [98, 99].

- Hybrid techniques. Similar to human processes, this approach uses both holistic and feature-based techniques.

Fully automated face recognition systems employ a face detection step before applying holistic and/or feature-based techniques (as in Figure 2.5). In the face detection step the aim is to determine automatically the presence of a face on still photographs or video [100]. Face detection techniques can be categorised into 3 groups based on: knowledge (algorithm has to follow rules such as distance of eyes) [101], template matching utilising detection of facial features [102], and invariant features such as skin colour segmentation [103, 104]. Some challenges in face detection are caused

due to the inherited diversity of faces (e.g. shape, texture, facial hair, ethnicities), head pose, facial expressions, occlusions (e.g. glasses) and scene complexity. Scene complexity factors include clutter caused due to environmental conditions (e.g. illumination variations, rain, snow, wind, shadows), and occlusion (face to face and face to scene occlusion) [81, 100]. Some of the face detection techniques are also applicable for the detection of other objects such as cars and humans [105]. Section 2.1.3 provides some more information on detection techniques.

This thesis investigates partial automated systems (automated face detection is not performed) utilising 3 basic and popular holistic techniques [51] with faces where both eyes are visible (not profile faces): a) Principal Component Analysis (PCA), b) Linear Discriminant Analysis (LDA), and c) Kernel Fisher Analysis (KFA). Facial regions are extracted based on manually obtained eye coordinates and later the faces are normalised in terms of geometry (i.e. orientation) and size (i.e. number of pixels). The following points summarise the 3 aforementioned techniques implemented by utilising an existing MATLAB Toolbox [51], in the experimental work in Chapter 5. PCA has been implemented utilising the algorithm by Turk and Pentlan [106], KFA utilising the algorithm by Liu [107] and LDA utilising the algorithm by Belhumeur [108].

- *PCA*. In case of greyscale (single-channel) images, the input to this method is a training dataset consisting of N number of facial images (each image of $k \times k$ pixel dimensions) as the example in Figure 2.6a. Later, each facial image is transformed to a single k^2 element vector. After the transformation of each individual image to the vector space, the result is the creation of a single matrix of the form $k^2_{rows} \times N_{columns}$. The $N_{columns}$ represent the individual face images that have now been transformed to vectors. PCA is applied to this resulting matrix. Firstly, the average looking face (see Figure 2.6b) is represented by the mean vector and it is subtracted from each of the image face vectors in the matrix (this step normalises the face vectors). Later, the covariance matrix is calculated. The eigenvectors are obtained from the covariance matrix and include most of the variance in the data. The outputted

eigenvectors would normally be of a reduced size matrix $k^2_{rows} \times m_{columns}$ and it would be representative of the initial training vector space dataset $k^2_{rows} \times N_{columns}$. The eigenvectors can be transformed back to images called eigenfaces (see Figure 2.6c) as the dimensionality of the image pixels k^2 is sustained throughout the process.

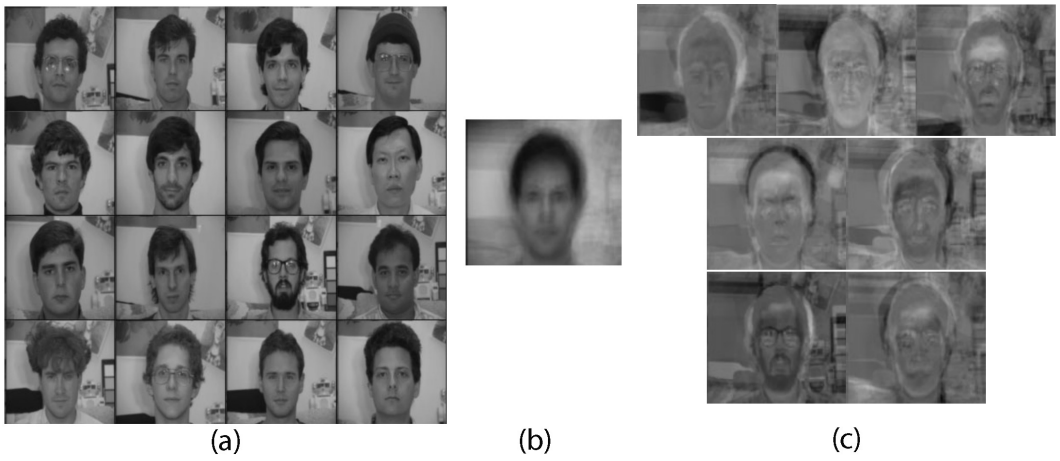


Figure 2.6: The process of obtaining the eigenfaces. Figure (a) represents the dataset of training facial images, (b) the average face, and (c) the obtained eigenfaces calculated from the input dataset in (a). Adapted from Turk *et. al.* (1991) and Jafri *et. al.* (2009) [85, 106].

As we can see in Figure 2.6 from an initial dataset of 16 faces, PCA has outputted 7 eigenfaces corresponding to the highest eigenvalues. The visualisation of eigenfaces (see Figure 2.6c) represents the most prominent deviations from the mean (see Figure 2.6b) in the dataset. For example, eigenfaces capture variations in the hair area, eyes and positioning of head area. In this example, variations in the background area have also be captured. Intervention of background information should be eliminated by either cropping the images to include only facial regions or be kept constant, otherwise it will affect the correct recognition results. This method compares an unknown image to a dataset of known images and if the background information is not kept constant or eliminated, it will become part of this comparison. When a new face is added to the dataset, the eigenfaces are recalculated.

Each face in the original dataset can be represented by a weighted sum of each of the eigenfaces plus the mean vector of the average face (i.e. the

result is known as a weighted vector). When an unknown face is given to a face recognition system, its weighted vector is calculated using the eigenfaces and the average face. Then, the unknown weighted vector (which represents the unknown face) is compared with the known weighted vectors (faces in the training dataset) using a distance measure (e.g. Euclidean distance). Recognition can be achieved empirically (i.e. by operators such as police officers) by specifying a threshold value to the distance measure. For example, to set the threshold value to first provide face images that have achieved 80% match with the query face image. If a correct match is not achieved then the threshold value can be decreased (e.g. to 70%) for more facial matches.

- *LDA*. This method follows the same principles as the PCA method. The main difference is that LDA utilises the relationships between the facial images in the training dataset and their relationship to the training dataset as a whole. For example, when the training dataset is created the facial images of a single individual are labelled in the same class (i.e. the aim is to minimise within-class variance) and the images of each individual are in different classes (i.e. the aim is to maximise between-class separation to increase discrimination). This is achieved by calculating a separation matrix in order to achieve a cluster separation analysis [109]. The resulting eigenvectors when transformed back to images are called fisherfaces. Figure 2.7 demonstrates the visual differences between eigenfaces and fisherfaces. Fisherfaces seem to not have included unrelated variations in the faces caused by lightness and head pose, whereas such variations are more visible with eigenfaces.

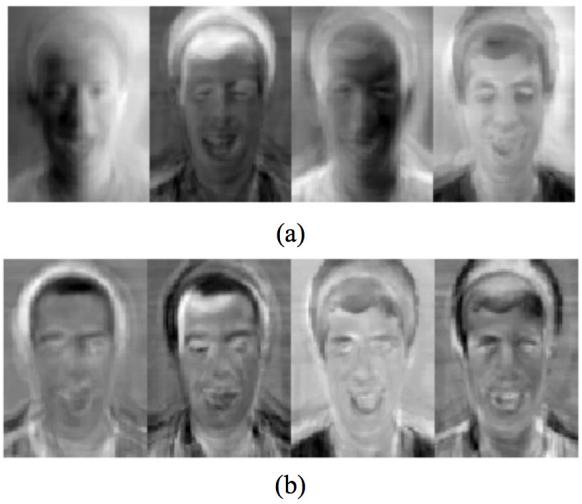


Figure 2.7: Eigenfaces vs. Fisherfaces . Figure (a) corresponds to images of eigenfaces and figure (b) to fisherfaces. Eigenfaces have captured variations caused from lighting and/or change of head pose. Whereas, fisherfaces seem to have decreased such unrelated variations from the face images. From Swets and Weng (1996) [110].

- *KFA*. Also known as Generalised Discriminant Analysis (GDA) is a kernelised version of LDA. Its basic idea is to map the input original sample vectors into a high dimensional feature space and then use LDA on that new feature space. The nolinearity enables the extraction of nonlinear discriminant features [111, 112].

Face recognition systems are aiming to be competent for real-world applications. A number of video/image face datasets and procedures are available for the assessment of the performance of face recognition systems [33, 38–40]. In most cases, face recognition systems are evaluated based on their performance of correct recognition from large datasets and individual properties of each facial image are not considered. For example, in a recent performance evaluation by the National Institute of Standards and Technology (NIST) [41] current state-of-the-art face recognition systems were assessed using web-camera images. These web-camera images were just labelled as being of poor quality and they were included to show how recognition accuracy degrades in uncontrollable conditions such as surveillance situations. Individual and specific scene content properties such as over or under exposed faces were not taken into consideration.

Furthermore, Aggarwal et al. [42] have identified that face recognition systems per-

formed differently on different datasets due to the dissimilar scene properties of the facial imagery under each dataset. For example, results of performance comparison between PCA and LDA techniques differ significantly among different databases and no conclusion can be made on the best performing one [43]. LDA seems to be handling variations in illumination better than PCA [66, 110]. These findings underline the scene-dependent nature of face recognition algorithms. Scene dependency can be overcome by characterisation and classification of scene properties (see Section 3.5), which will allow specific scene content characteristics (e.g. an under exposed or a low contrast scene) to be taken into consideration when analysing results. Knowing exactly with what content characteristics a system fails can contribute to further improvement of such systems.

The findings in automated face recognition using still imagery with JPEG [31] and JPEG2000 [32] standards agree that compression does not adversely affect automated face recognition performance [33–37]. In a study by the Face Recognition Vendor Test (FRVT) with JPEG compression [33], the findings have shown no deteriorated performance with images compressed to 0.2bpp [33] and performance deteriorated under 0.2bpp. The same study has even reported an increase in performance of face recognition systems with compressed images to 0.8bpp and 0.4bpp [33]. These results are consistent with another investigation conducted by Delac and co-authors [35, 37]. The compression amount of 0.2bpp in still photography, is equivalent to a compression bitrate in video of around 208kbps (kilobits per second) for a full D1 PAL resolution (720×576) at 25 frames per second (fps). Bitrate denotes the number of bits conveyed/stored or processed/transmitted per unit of time (see Section 3.1).

In one recognition study [113], the performance subspace techniques (including LDA and PCA) was assessed with JPEG and JPEG2000 compression standards. The probe facial images were compressed and the training/gallery images were in an uncompressed format. The amount of compression ranged from 1bpp (light compression) to 0.2bpp (high compression). The performance of the automated face

recognition algorithms was assessed using the rank one recognition rate [38] identifying if the top match is correct. The investigators utilised 3 main categories of probe images: i) different facial expressions, ii) different illumination conditions, and iii) images taken at different points in time (i.e. to study the ageing effect). The results have shown few correct recognition increases when the images were lightly compressed but none of these increases have been found to be statistically significant. Overall, they have found that higher compression (0.5, 0.3 and sometimes even 0.2bpp) is more suitable for recognition with datasets conveying different expressions and illumination conditions. Whereas, lighter compression is more appropriate for recognition of images taken at different points in time (age). Additionally, they have found that the application of JPEG2000 standard had less of an effect in the recognition rate than the JPEG standard.

The aforesaid increases in performance of face recognition algorithms with compression might have occurred due to the ability of compression algorithms to sharpen edges when adopted in small amounts. For instance, Ford [114] investigated the behaviour of JPEG and fractal compression standards in relation to contrast. He has found an increase of contrast when compressing at around 0.44bpp. His contrast method included the capture of an edge and calculations of contrast ratios (known as microscopic contrast for sharpness measurements, see Section 3.3.3) for both the uncompressed original and its compressed versions. Both JPEG and fractal compressions produced increase in contrast (JPEG 10% increase and fractal 1%). JPEG compression derived a greater increase in contrast as the ringing artefact (i.e. Gibb's phenomenon) is occurring by either side of an edge resulting in a localised increase on an edge and thus image contrast [115]. This is not the case for fractal compression as the ringing artefact occurs only on one side of an edge.Furthermore, Ford [114]] has found that the increase of compression amount results in reduction of intensity levels (i.e. tone levels reduction).

Age, illumination and pose variations are the most challenging situations affecting the performance of face recognition systems [36, 116]. Adler and Dembinsky [44]

have investigated the relationship between human evaluators and face recognition algorithms in relation to biometric image quality. Human evaluators had a strong correlation with each other as were the face recognition algorithms. However, they have found that derived biometric image quality scores do not correlate between human perception and automated algorithms. Figure 2.8, presents an example of facial images that have scored the best and worst in terms of image quality between human evaluators and automated algorithms. No differences in relation to image quality can be observed between humans and algorithms for the best category images. For example, all images seem to be correctly exposed (see Figure 2.8). On the other hand, some differences can be observed for the worst category images. For example, the very dark image under the worst category for humans is not included under the worst category for algorithms. Also, the algorithms have located 2 over exposed faces in the worst category that are not included under the worst category for humans.

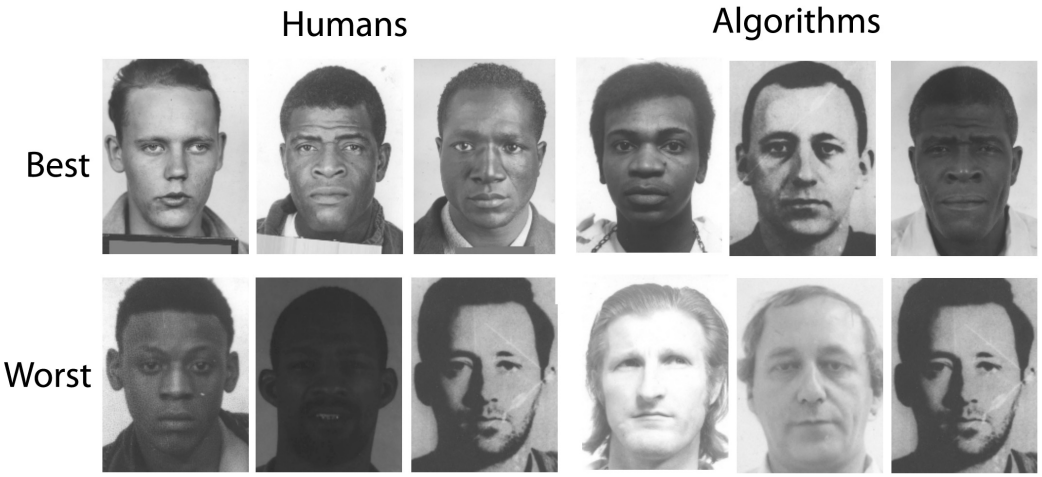


Figure 2.8: An example of facial images presenting the best and worst image quality scores for human evaluators and automated face recognition systems. From Adler and Dembinsky [44].

2.1.3 Video analytics with the sterile zone scenario

Video analytics (VA) are computerised autonomous systems that analyse events from camera views for a variety of real-world applications. For example, applications relating to video surveillance, retail and transport industries [26,117]. One example

of a retail application is the counting of people entering a store for statistical and marketing purposes [118]. Generally, VA systems take as an input a visual scene and produce as an output meaningful quantitative data and/or an alarm response. For example, VA systems produce numbers in the people counting application and alarms in an intruder detection application. VA systems are expected to work: a) in uncontrollable environmental conditions, b) with unknown camera qualities, c) continuously for long hours, and c) with little or no human interaction [25]. Such systems have to adapt and learn from all the aforementioned variables in order to provide meaningful outputs.

VA systems can operate in real time (i.e. incidence alert) and in post event analysis (i.e. when incorporated within a recorder for event-based retrieval) [119]. VA systems are automated tools that the police utilises to complete recognition tasks from CCTV footage. In consideration of the vast amount of video CCTV data [53, 54], the monotonous task of human visual examination of video data and the effective impact that CCTV has on conviction of crimes [2], automated systems are a beneficial tool to the police and perhaps the future of policing.

Some examples of surveillance applications include detection and reporting of: license number plates, abandoned objects, vehicle counting, single or multiple people [120–122]. One of the few surveillance applications able for autonomous analysis [26, 27] is the sterile zone (SZ) scenario from the iLIDS dataset [50], which will be investigated in this thesis.

The SZ is a low complexity scenario, consisting of a fence (not to be trespassed) and an area with grass (see Figure 2.9). The VA system needs to alarm when there is an intruder entering the scene (an attack). The task of the SZ scenario is human detection.

The current available machine vision techniques for the analysis of video content are overwhelming in volume. Figure 2.10 illustrates a general work-flow of a video analytics system. In reality, manufacturers of such systems might include a pre-

Figure 2.9: The Sterile Zone scenario from the iLIDS dataset [50]. From left to right, the camera to subject distance is far, medium and close.

processing stage before segmentation. This pre-processing stage might include actions such as frame format conversion, noise removal, decompression and object enhancement (this is applicable also to face recognition systems) [123, 124]. Additionally, each stage in the work-flow in Figure 2.10 could consist of an integrated set of techniques.

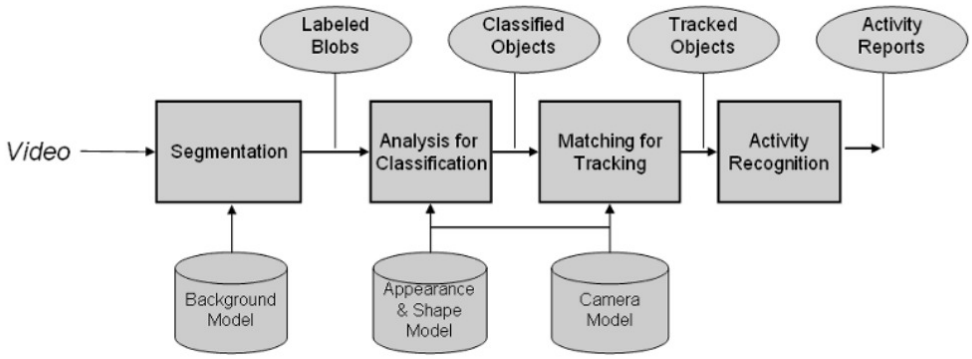


Figure 2.10: Video analytics software components. From Gagvani (2009) [25].

The following points summarise the stages of the work-flow in Figure 2.10:

- *Segmentation.* The segmentation stage normally assumes a static camera and separates foreground pixels from background pixels. Background pixels are subtracted. Further, foreground pixels are grouped into blobs [125, 126]. Blobs are connected sets of pixels. Blobs created due to rain, snow, wind, water, shadows, reflections are called clutter and can be removed by applying clutter removal techniques [127]. Results from the segmentation stage are labelled blobs and correspond to unique objects within the scene. Segmentation performance is affected by clutter (due to illumination variations and

environmental changes), contrast across the scene (due to time of day/season and street lighting), low resolutions (caused by wide field lenses employed in the surveillance industry) and perhaps compression (because of loss of high frequency information and creation of artefacts) [25]. Compression standards are developed around the sensitivity of the human eye in order to make compression artefacts less or not visible to humans. Nevertheless, these 'invisible' artefacts might affect the performance of mathematical algorithms applied by VA systems. Segmentation is under the object detection category. Paul *et. al.* [128] present some available human detection techniques for surveillance applications such as background subtraction, optical flow and spatio-temporal filtering.

- *Analysis for classification.* This stage assigns a class label to each segmented blob such as person or vehicle. The class label depends on the presence of persistent and discriminant features over multiple frames. Tsuchiya and Fujiyoshi [129] investigated the importance of various features for the classification of objects such as a vehicle, a single human, a human group and a bike. The same researchers have proposed shape-based features, texture-based features, and motion-based features that might be reliable for classification of the aforesaid objects. Their proposed features are invariant to various changes caused by environment, scaling, viewpoint and lighting. Viola and Jones [130] have utilised a supervised learning approach using a large dataset of visual features. In case of number plates, a recognition stage is applied (same as face recognition) in order to match the captured number plate in the scene with a database of registered number plates. Automated face recognition is not considered to be within the realm of video analytics as some control over illumination conditions and camera set up is commonly practised (automated face recognition performs the best under controlled conditions). In general, classification works well on distinguishing between humans and vehicles but problems arise with more complex scenes when multiclass classification is required. Other issues include the poor extraction of features due

to segmentation/object detection errors [25].

- *Matching for tracking.* This stage deals with the tracking of individual blobs within successive frames by locating their position. Later, it contributes to behaviour or activity interpretation. Numerous approaches to tracking exist depending on the context/environment of the scene [131]. Some of them are based on points (objects within frames are represented by points), and/or kernels (referring to the object shape and appearance), and/or silhouettes (estimating the object region within each frame). Tracking techniques face many challenges as blobs might reveal different features within successive frames. For example, a person moving in the scene might change pose and orientation. Some other challenges are occlusions (e.g. a person walking behind a van or 2 people occluding one another) and frame rate reduction. Low frame rate is considered equivalent to 'abrupt motion', or discontinuity by tracking techniques [132]. Tracking techniques frequently use motion continuity and their performance is consequently affected by low frame rate [133, 134]. In Europe the standard video frame rate for television is 25 frames per second (fps) (or 50i interlaced fields). Commonly, security systems record/transmit video data at lower frame rates in order to satisfy storage and transmission requirements.

- *Activity recognition.* Once blobs have been tracked then their motion can be described with respect to the rest of the scene (e.g. activity of other blobs and background). There are numerous models available for activity analysis and recognition such as track trajectories and ground plane, which describe activities such as fallen person, slow moving vehicles [135, 136]. Other approaches include kinetic models for human activities (e.g. crouching, bending and jumping) [137].

The extended number of available video content analysis techniques is a proof of the complicated nature of video analytics systems. They do not consist only of few techniques but rather an integrated variety of techniques which are all inter-

connected. This makes the testing of such systems difficult in terms of drawing conclusions for the appropriateness of each utilised technique integrated in a complete video analytics system (this is also applicable for face recognition systems). Each individual technique has limitations. For example, feature-based detection and tracking techniques have problems with feature visibility, scale changes and low contrast. In comparison, motion-based techniques face problems with differentiating fake motion from object motion, clutter, object to object and object to scene occlusion [138]. Thus, it is really important that the performance of such systems is evaluated with representative datasets of specific applications. The following Table 2.1 summarises the available datasets and benchmarking processes for system performance evaluation. Performance results use ground truth data (e.g. information on exact timing of an individual in the scene) and distance metrics (i.e. from the ground truth information) [139].

Benchmark	**People**	**Vehicle**	**Animals**	**Objects**
PETS [140]	✓			✓
i - LIDS [50]	✓	✓	✓	✓
CAVIAR [141]	✓			✓
VACE [142]	✓	✓		✓
TRECvid [143]	✓			✓
Daimlerchrysler [144]	✓	✓		
PASCAL [145]	✓	✓	✓	✓

Table 2.1: Available benchmark datasets for the evaluation of video analytics algorithms. Adapted from Tawiah (2010) [138].

Each of the datasets in Table 2.1 use their own distance or performance metrics. Those metrics can be found in the references provided next to the name of each dataset. In conclusion, performance evaluations either for a specific algorithmic technique (e.g. tracking algorithm) or a complete video analytics systems (i.e. with all the components/stages) is based on overall performance using large amounts of footage and individual scene content properties are not taken into consideration. For instance, often a ROC (Receiver Operating Characteristic) curve is used to visualise results. ROC curves plot true positive rate (correct positive results) against false positive rate (incorrect positive results)(see Figures 2.11 and 2.12).

Little research has been done in the area of image compression and video analytics systems, because currently only few scenarios are capable for autonomous analysis [26]. However, this area is receiving a large amount of research investment, even though it is still in its infancy [26]. In a world of rapid technological change, video analytics will need to be more flexible and be suitable for use in post-event forensics and with limited transmission bandwidth (e.g. through an Internet Protocol network).

In one investigation [46] with the SZ scenario (or intruder detection) and H.264 / MPEG-4 AVC video coding standard, the results have shown the performance of the video analytics system to be affected at 220kbps, either by not detecting an attack, or producing a slower alarm response time. That work investigated 11 attacks with only one VA system. In this thesis, a much greater amount of footage (or attacks) and number of VA systems are investigated.

Another investigation by Poppe *et. al.* [45] has pointed out to the lack of research in the area of video analytics systems with compressed footage. The same authors have tested a background subtraction technique called Gaussian Mixture Models (GMM) with the H.264/MPEG-4 AVC video coding standard applied using Constant Bit Rate (CBR). CBR keeps the bitrate constant and varies other parameters (e.g. quantisation parameter (QP)) based on scene content. Part of their results is presented by the graphs in Figures 2.11 and 2.12, which correspond to different scenarios. The performance of GMM was not only dependent on the amount of compression (i.e. the scene in Figure 2.11 has been affected less by compression than the scene in Figure 2.12) but also on scene content (i.e. difference overall performance when comparing the originals of each scene). This is indicative of the scene-dependent nature of computational algorithms.

In conclusion very low bitrates degrade performance (Figure 2.11) whereas high bitrates have a positive effect on the performance (Figure 2.12). This might happen because compression can sometime act as a noise filter and thus make the visual scene simpler for an algorithm. Also, at low amounts of compression (high

bitrates) it has been previously reported that compression methods have the ability to sharpen edges [114]. Again, the experiments by Poppe *et. al.* include only 2 video scenes and a greater number of scenes would be required in order to understand the scene dependency phenomenon in more detail.

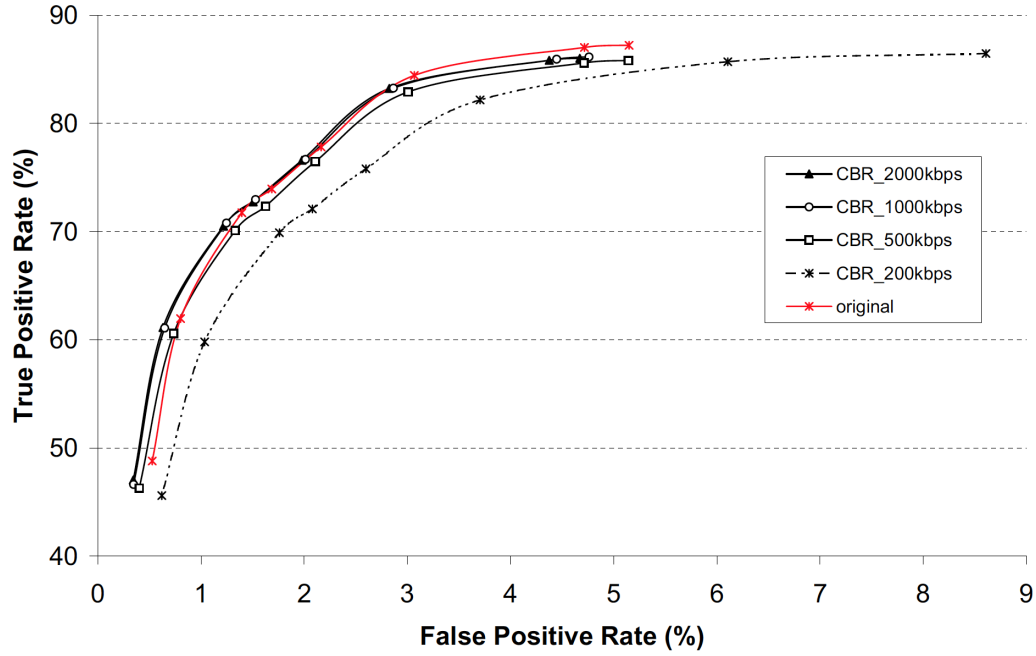


Figure 2.11: GMM performance with compression and the PetsD2TeC2 sequence (384x288). From Poppe *et al.* (2009) [45].

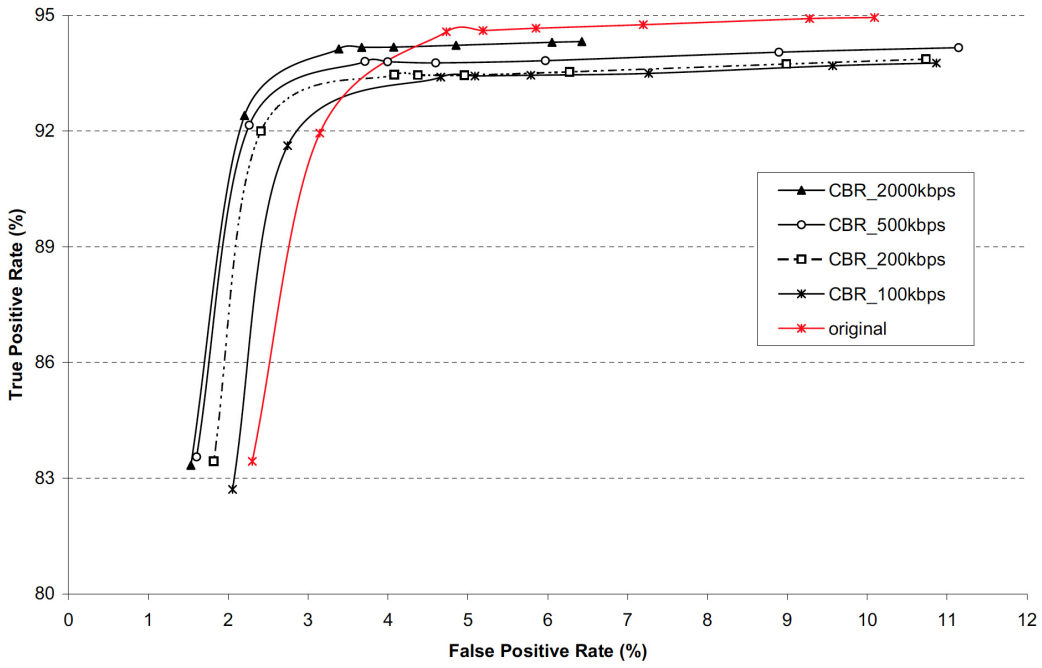


Figure 2.12: GMM performance with compression and the Indoor sequence (320x240). From Poppe *et al.* (2009) [45].

2.2 Discussion

Image quality for human recognition tasks relates to the visibility of useful information. The capture of the useful information will be affected by the unconstrained conditions commonly encountered by CCTV systems (e.g. illumination variations and compression). The unconstrained conditions will affect human and automated systems differently. The different behaviour between human and automated systems will become apparent in Chapters 5 and 6. For instance, under exposed (low facial lightness) scenes are affected by compression less than correctly exposed (medium facial lightness) scenes for automated face recognition systems. Whereas, for the human system the correctly exposed scenes are affected less by compression and the under exposed scenes the most.

Adler and Dembinsky [44] have also identified the different behaviours among automated and human systems for face recognition tasks (see Figure 2.8). Furthermore, Korshunov and Ooi [146] have identified that surveillance automated systems (face detection, recognition and tracking) accept significantly more compression compared to humans and have also pointed out to the need for alternative image quality measures suitable for automated systems. Understanding how automated systems behave with a variety of scene contents will contribute in identifying appropriate image quality metrics. Furthermore, in Section 3.2.3, background research proves that compression performance is also influenced by scene content properties. This phenomenon is known as scene dependency. Scene dependency can be overcome by characterisation and classification of scene content properties/characteristics (see Section 3.5.1)

When the performance of a visual system (both for automated and human) is assessed then, a number of factors need to be understood before applying a testing methodology. Chapter 3 expands more on the factors connected to the aspects of image quality and compression.

CHAPTER 3

Video compression and image quality for CCTV imagery

This chapter presents the aspects of image quality and video compression that may affect the completion of police tasks from Closed-Circuit Television (CCTV) footage. Visual police systems, either automated (i.e. automated face recognition) or human operatives, can be assessed with compression using controlled footage in terms of image quality attributes and scene content properties. CCTV imagery is used for the completion of police tasks, which have been described in Chapter 2. The following sections provide information on video compression for both 'standard' and CCTV industries. Additionally, the definitions of image quality are discussed together with physical image quality attributes (i.e. attributes that can be measured objectively and also subjectively such as tone reproduction) and psychological image quality attributes (i.e. attributes that are only measured subjectively, such as image usefulness). The methods of collection and quantification of observers' responses in psychophysical investigations are explained, together with the factors that affect such procedures. Information on scene content characterisation and scene classification is provided.

3.1 Video fundamentals

Video is a presentation of visual imagery using sequential still images called frames. When this visual imagery is presented between 24 and 60 frames per second (fps), the illusion of motion is created [147]. Video frames use either progressive, or interlaced scanning. Scanning in this context refers to the way a system displays, stores, or transmits video data. For example, LCD (Liquid Crystal Display) computer monitors use progressive scanning and CRT (Cathode Ray Tube) television displays use interlaced scanning. In progressive scanning a frame conveys the complete set of even and odd lines of an image (see left hand side illustration in Figure 3.1). When the video uses interlaced scanning, a frame consists of 2 fields captured at different times in a successive order (see right hand side illustration in Figure 3.1). One field conveys the even lines and the next one the odd lines. Although, the transmission signal for a standard definition (SD) video field is between 200 to 300 lines, the final display is between 500 to 600 lines [148]. Moreover, a progressive scan video signal consists of twice the amount of information from an interlaced scan video signal. Figure 3.2, illustrates an example of how interlaced video appears on an LCD. The interlace effect on the LCD distorts the visual information.

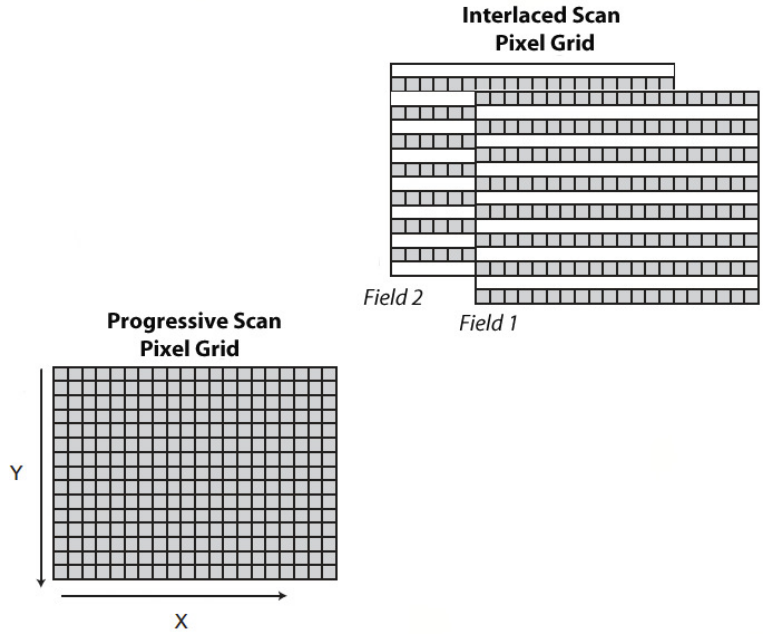


Figure 3.1: Progressive and interlaced scanning. From Austerberry (2005) [148].

Figure 3.2: The interlace effect. The image on the left illustrates the interlaced effect on an LCD display and the image on the right how the image appears when 1 of the fields is removed.

At the acquisition stage, the camera produces signals of the primary colours red, green and blue (RGB). These signals are further processed for transmission and storage. About 3 main colour systems have been standardised in order to maintain compatibility for transmission and storage. These are Phase Alternating Line (PAL), National Television System Committee (NTSC) and Sequentiel Couleur á Memoire (SECAM) [149].

In many European countries, including the United Kingdom (UK), PAL colour encoding system is used to broadcast television at 720×576 interlaced lines and 50 fields per second or 25fps (frames per second). PAL carries Y' U' V' components, where Y' is the luminance and U' and V' carry the chromatic information. The luminance bandwidth in the UK is 5.5 MHz (megahertz) and for each colour component the bandwidth is 1.5 MHz. This difference in bandwidths is due to the greater sensitivity of the human eye to luminance changes compared with chromaticity [150].

The digitisation process of the analogue video signal involves filtering, sampling and quantisation. Filtering is used to eliminate aliasing artefacts prior to sampling. The process of sampling is the conversion of the continuous signal to a discrete signal (e.g. to horizontal and vertical image coordinates). Temporal sampling is used as a third dimension, in addition to horizontal and vertical sampling (see Figure 3.3). Quantisation involves mapping the amplitude of the sample values to a smaller set

of points. The ITU Rec. BT 601 specifies that standard definition television frames consist of 720 lines of non-square pixels. This is normally referred to as the pixel aspect ratio, which it is not square for video.

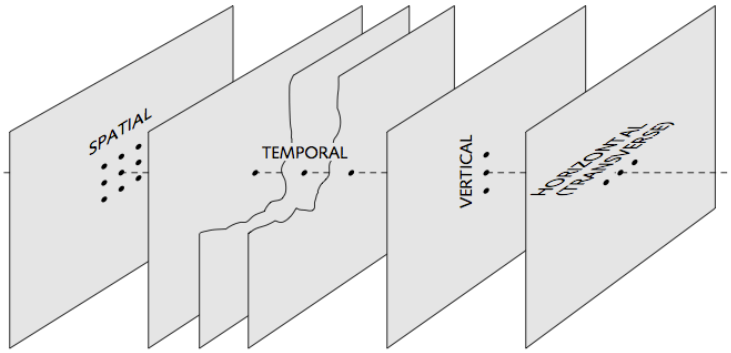


Figure 3.3: Spatio-temporal domains of video. Sampling in 3 axes: horizontal, vertical and temporal. From Poynton (2003) [147].

Storage of 'uncompressed' video data demands a considerable amount of space. For instance, a recording of PAL 'uncompressed' 8-bit footage for 5 minutes could occupy 1.22GB (Gigabytes) of storage. Bitrate refers to the number of bits conveyed/stored or processed/transmitted per unit of time. CCTV systems record / transmit video data continuously for days and as a consequence the original 'uncompressed' data are subjected to compression in order to enable efficient storage and transmission. Video compression is used to reduce data that are produced by video sources (e.g. video cameras). The methods utilised for data reduction include frame rate reduction and re-coding of the initial information within a video signal. The latter is referred to as video compression.

In Europe the standard frame rate for television is 25fps, although security systems commonly record at 5fps and below in order to reduce storage requirements. However, reducing the standard frame rate increases the possibility of missing important information from the initial video sequence. Essentially, if a suspect's face appears in 3 frames within a second when recording at 25fps, the reduction in the recording frame rate to 1fps will result in a significant reduction in the probability of obtaining 1 of the 3 frames that include the face. Additionally, as it has been mentioned in Section 2.1.3, low frame rate is considered equivalent to 'abrupt motion', or discontinuity by tracking techniques often incorporated within automated

systems (face recognition and automated systems) [132–134].

Video compression in the security industry employs proprietary formats based on industry standard compression algorithms. There are 2 main ways in which data are compressed in the CCTV industry: a) compression on storage (e.g. Digital Video Recorder-DVR), and b) compression on transmission (e.g. Internet Protocol-IP cameras). Currently, the most popular proprietary compression algorithms in the security industry are based on the H.264/AVC and JPEG standards for both transmission and storage [21–23].

Low bitrates are favoured in the CCTV industry, since they allow more hours of recording and lower the cost of the system/transmission. However, they compromise the usefulness of the recorded imagery. The H.264/AVC compression algorithm was chosen for investigation in this thesis. Section 3.2.1 provides a brief description of the H.264/AVC standard.

ITU-T and ISO / IEC JTC 1 have produced joint video standards (such as H.264 / AVC), which specify only the decoding part in order to ensure interoperability and syntax capability between different technologies implementing the standard. This allows developers of compression algorithms to optimise compression implementations based on specific applications, to trade-off costs, efficiency in terms of image quality, speed, error resilience and hardware requirements. Image quality is not specified in the standards and different implementations of the same encoder will produce different 'compressed qualities'.

3.2 Video compression

Video compression algorithms use techniques, such as predictive coding, to exploit the correlation of the video signal between neighbouring pixels and successive frames [151–153]. Predictive coding is used for both lossless and lossy compression, which are the 2 types of compression [31, 154]. The ultimate aim of any compression algorithm is to represent the initial captured visual information using less

data.

When data is subjected to lossless compression, the data that represent the sequence of images (video) are reduced without any loss of information [155], whereas in lossy compression information is lost. Lossy compression will be the focus of this thesis, since it is the one used in the security industry for storage and transmission purposes. Also, with lossy compression very low bitrates can be achieved whilst this is not the case for lossless compression. Lossy compression removes mainly invisible information at lighter compression levels [156]. The removal of this invisible information will not have a consequence to human observers but might have a consequence to automated algorithms. As compression levels increase, it starts removing visible information.

Lossy compression is a distortion process that can potentially affect the visible information in video [157] and as a consequence the image usefulness of the CCTV footage to complete a police task. In video compression, 2 main categories of techniques are used. These are the predictive coding and transform coding techniques, which are described briefly by the following paragraphs.

- Predictive coding techniques: When predictive coding (or hybrid predictive coding) is applied, decorrelation occurs both spatially (within an individual frame known as intra frame predictive coding) and temporarily (within successive frames known as inter frame predictive coding). Predictive coding techniques assume correlation of pixels within individual and successive frames (i.e. cross correlation) [158]. Commonly, the changes in video content occur due to changes of objects motion. The background often stays the same within successive frames. This is used in motion compensation (MC) techniques in order to achieve high compression gains [153, 159]. Other motion compensation techniques include: global motion compensation (prediction for global motion on the entire frame caused by panning, titling and zooming) [160], stripe panoramic (the content of the background is represented by a single still image in the sequence, which is transmitted separately from the

foreground object) [158], segmentation (image is separated into segments of coherent regions, textures and objects, which are later coded and prediction occurs on the motion of a segment between successive frames) [153, 161] and semantic segmentation (object based segmentation for example people and the use of 2D/3D object models for the coding) [162, 163].

- Transform Coding techniques: There are 2 main methods under this category, the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT). The DCT divides an image into (usually) 8×8 pixel blocks that are later coded individually; the method introduces quantisation artefacts (see Section 3.2.2). DCT is used by lossy compression algorithms such as JPEG, MPEG, and ITU standards. The DWT technique divides an image into low pass and high pass components that are later filtered using wave functions of different shapes, producing smoother and fuzzier compressed images [158,164].

Often, video compression techniques are used in combination. For example MPEG and ITU-T video standards are known for employing hybrid techniques such as block-based MC/DCT techniques [153, 159]. Quantification of the amount of data reduction obtained from applying a compression algorithm is often represented by the compression ratio. The compression ratio is simply a ratio of the uncompressed size (or data rate) to the compressed size (or data rate).

The aforementioned techniques have been described in order to emphasise the complicated nature of video compression algorithms. The following Section 3.2.1 provides a more specific explanation of the H.264/AVC compression algorithm that has been employed in the experimental part of this thesis.

3.2.1 The H.264/AVC Compression Standard

H.264 / AVC is also known as MPEG-4 part 10, or Advanced Video Coding (AVC). The H.264/AVC compression standard is the output of the collaboration between the International Organization for Standardisation's MPEG group (ISO/IEC JTC

1 / SC29 / WG11) and the International Telecommunications Union's video coding experts group (VCEG, ITU-T / SG16/Q.6) [24]. H.264 / AVC is a hybrid video encoder, exploiting both spatial and temporal redundancy and uses a 4×4 integer transfer (an approximation of the 4×4 DCT). H.264 / AVC produces blocking artefacts that become more visible at low bitrates.

Figures 3.4 and 3.5, illustrate the basic encoding and decoding hybrid structure correspondingly of H.264/AVC. In the encoding part, the first frame of the sequence is split into macroblocks and is coded in intra mode with the use of spatial predictions. For the rest of the successive frames, inter frame coding techniques are used. The residual difference between the original and its prediction is transformed by a frequency transform. The coefficients of the transform are later scaled, quantised, entropy coded and finally transmitted together with the predictions [165].

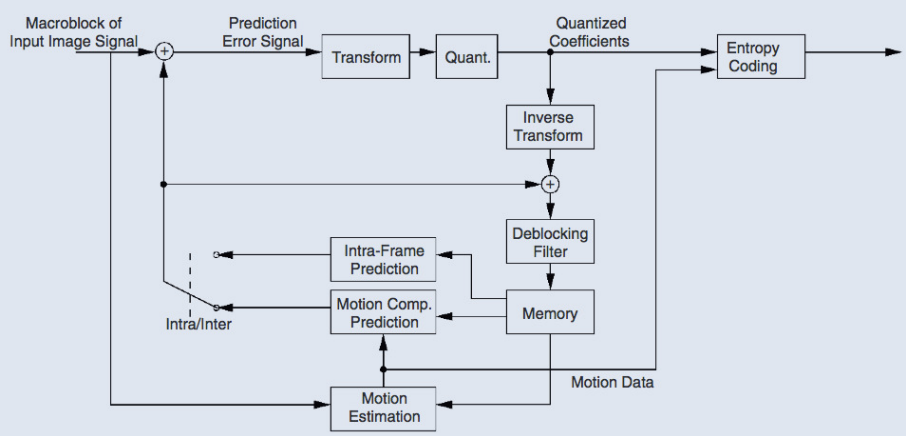


Figure 3.4: General encoding structure of H.264/AVC. From Ostermann *et al.* (2004) [166].

In the decoder part, the quantised transform coefficients are inversed-transformed and added to the predictions. After deblocking filtering, the output is a reconstructed macroblock. Macroblocs are stored in a memory in a raster scan order (ordering of pixels by rows) [166] in order to allow prediction of subsequent encoded macroblocks.

Implementation encoders such as Joint Model (JM) and the Fast Forward Moving Pictures Expert Group (FFpmeg) are verification models used for compliance testing of 'industrial' implementations. These are often used by the scientific commu-

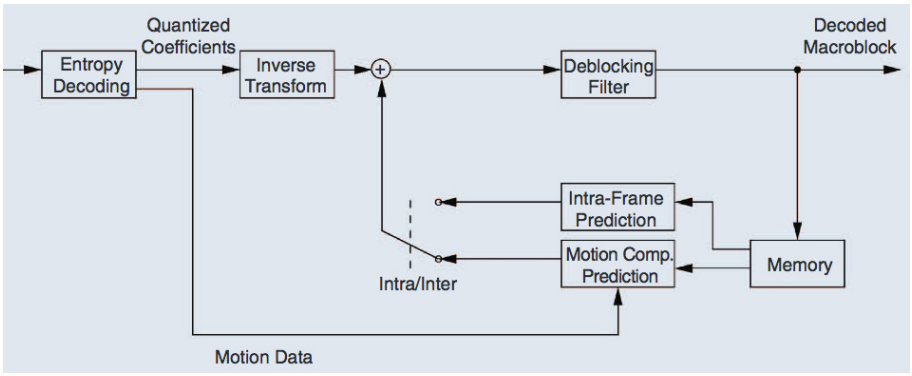


Figure 3.5: General decoding structure of H.264/AVC. From Ostermann *et al.* [166].

nity; they allow the setting of over 50 parameters, such as quantisation parameters, I (i.e. inter frame), P (i.e. predicted frame) and B (i.e. bidirectionally predicted frame) frames and the target bitrate. These verification models tend to apply 'high quality' compressions whilst encoders in the consumer and CCTV industry apply 'low quality' compressions.

3.2.2 Compression artefacts

Compression techniques rely on the 'continuity' of a signal. For instance, they use information from neighbouring pixels and successive frames. When the signal is added with an artefact, such as noise (which might come from sensor, transducer and analogue to digital processes), then the compressor will require higher bitrates (less compression) and bit depth to compensate for that 'discontinuation' [157].

There are several types of artefacts, depending on the degree of compression. The most recurrent artefacts, in particular for compression algorithms using the discrete cosine transform (DCT) technique, are listed below:

- Blocking artefacts. The block and/or macroblock structures become more visible as compression ratios increase. Blocking is more visible in flat areas within imagery, as the visual human system is very sensitive to small brightness changes. Nevertheless, blocks tend to affect high spatial frequency image information (busy areas). Furthermore, blocking will be more noticeable when

a sequence consists of rapidly moving objects. For instance, intra frame coding might result in trails left behind and the object might be moving over blocked patterns (see Figure 3.6) [167].

- Decreased colour bit depth and colour bleeding. A smaller palette of colours will be available for higher compression ratios. Colour bleeding is caused by the suppression of high spatial frequency information and it is more dominant for wavelet-based compression algorithms such as JPEG2000.

- Mosquito noise or ringing. The high spatial frequencies on sharp edges are quantised more coarsely than the low frequencies. After quantisation the edges appear to have a pattern of blurred dots (see Figure 3.6).

Figure 3.6: Example of blocking and mosquito artefacts. The blocking artefacts are more visible on the flat areas and mosquito artefacts around the edges of objects within the image.

3.2.3 Factors affecting compression performance

The following factors have been identified to influence compression performance:

1. Content of the scene. Compression performance is highly dependent on scene content [29, 30, 47, 48]. In video, scenes with different motion properties (temporal differences), regions (spatial differences) and combinations of different spatio-temporal properties will require different bit-budgets leading to differ-

ent levels of compression. Figure 3.7 provides an example of the H.264/AVC encoder performance under different illumination conditions. The bright (at the bottom) and dark (at the top) scenes are more sensitive to compression in comparison to the correctly exposed scene in the middle.

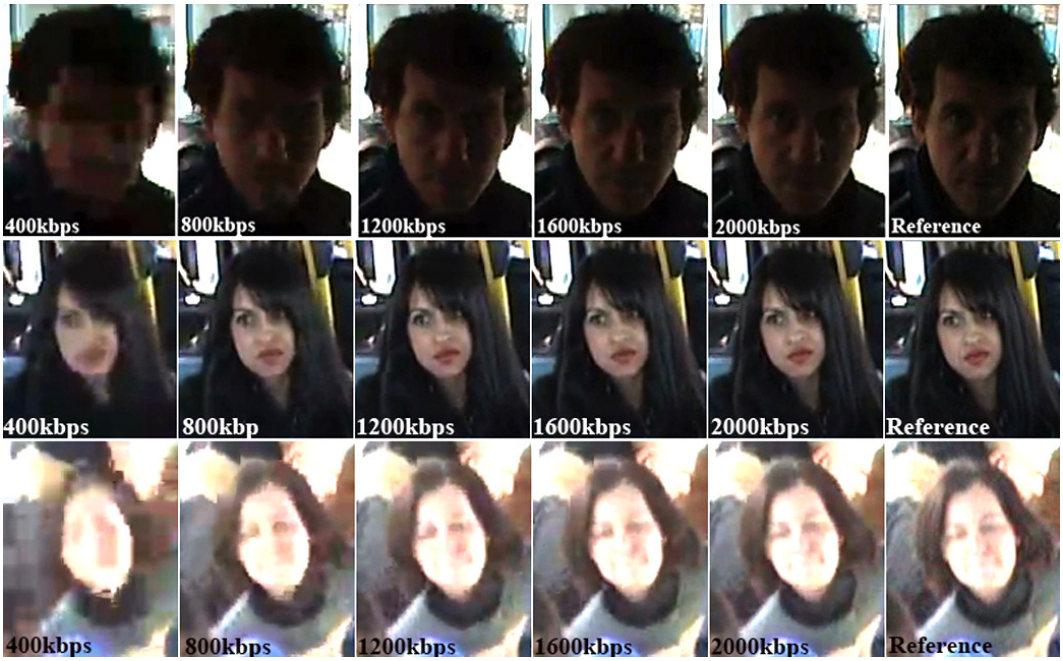


Figure 3.7: Example of compressed facial images. As the bitrates decrease the useful information decreases. Defining acceptable compression ratios depends on the original image and observers' acceptability standards. For instance, the bright (over exposed scene at the bottom) and dark (under exposed scene at the top) scenes are more susceptible to compression. They require lighter compression to achieve 'acceptable' responses from observers, in comparison to the correctly exposed scene (in the middle).

2. Compression Algorithm, its Properties and Settings. This can be whether the compression algorithm is based on Discrete Cosine Transform (DCT) or Discrete Wavelet Transform (DWT). DCT and DWT produce different types of artefacts, which appear at different areas (regions) within the imagery [48, 168]. Other settings and properties include intra-coding or inter-coding, quantisation parameters, reference frame selection ad the use of post-processing tools.

3.3 Image Quality Definitions

Image quality is a multidimensional and multidisciplinary domain, where (amongst others) observer judgements are used in the design and evaluation of imaging systems [169]. Images have different applications. Images that are meant for consumer purposes prioritise quality in terms of aesthetics (i.e. to satisfy visually the observers), while those used in scientific and industrial applications are aimed for information extraction and completion of tasks. In security, images are used for recognition and identification tasks.

Image quality, in its strict definition, is considered as being subjective since humans are the ultimate judges and end users of images. A number of well-known imaging scientists have defined image quality. Engeldum [170] has defined image quality as "the integrated set of perceptions of the overall degree of excellence of an image", Jacobson as "the subjective impression formed in the mind of the observer of the degree of excellence exhibited by an image" and Triantaphillidou [171] as "the subjective impression of goodness the image conveys". All these descriptions emphasise the importance of the observers' perception about the degree of subjective value of an image.

Perceptual image quality will be affected by the observers' memory and expectations (e.g. of colour reproduction); preferences (in terms of contrast, colourfulness and sharpness/blurriness) and cultural background. For instance, in one investigation the observers preferred a slightly higher average chroma than the unprocessed original [172]. Furthermore, research has shown that perceptual image quality is strongly correlated with naturalness, meaning that image naturalness is an important image attribute that influences perceived image quality [172–174]. Perceived image naturalness is influenced by the expectations and experiences the observers have in the world [175]. In this context, naturalness does not correspond to the exact representation of a scene but rather the memorised realities by the observers [172]. These memorised realities will differ among different observers. For example, it has been shown that people have strong positive associations with the colour of their

national flag [176]. Also, meanings and perceptions of colours differ among different cultures [177].

Image quality can be measured by applying both objective (quality metrics) and subjective (i.e. human investigations) approaches. Most desirable, objective approaches should agree with results obtained from subjective approaches. Subjective approaches are described in detail in Section 3.4. The following sections present a combination of subjective and objective approaches.

Moreover, the term image quality is used loosely often by non-experts to describe other image aspects that are related to image quality. For example, *Image distortion*, *image fidelity* and *image quality* are unique factors that evaluate different properties of images and imaging systems (see Section 3.3.1).

3.3.1 Distortion, fidelity and quality

Image distortion, *image fidelity* and *image quality* are important factors for the evaluation of images and imaging systems. In image distortion measures, the original (distortion free) image is normally available. Image distortion measures provide numerical differences between an original and a reproduced, or processed image. For example, distortion is measured by taking differences in pixel values. Distortion metrics, such as the mean square error (MSE) [178], root mean square error (RMSE), and signal-to-noise-ratio (SNR) [179], are often used to assess the effects of image processing methods (e.g. compression). More specialised distortion metrics exist for the measurement of colorimetric distortion, such as ΔE^*_{ab} that can be derived from the CIELAB colour space (i.e. see Section 3.3.4) and ΔE^*_{00} (i.e. this is a CIEDE2000 formula based on the CIELAB colour space) [180]. The aforesaid distortion metrics do not correlate with subjective results as they fail to predict subjective image quality across images with varying content properties such as edges, textured regions and luminance variations (i.e. the spatial structure of the image is not considered) [181–183].

However, distortion methods are still used to assess the level of information loss from image processes and more specifically for compression. For instance, peak-signal-to-noise-ratio (PSNR) is still often used in video compression in order to determine the level of compression for different scene contents [184,185]. Figure 3.8 illustrates an example of PSNR plotted against the bitrate (in kilobits per second) of a scene with a face (the face is in motion) on a uniform static background. As it is expected, the PSNR is larger when only the face area is considered alone than when the entire image/frame is considered.

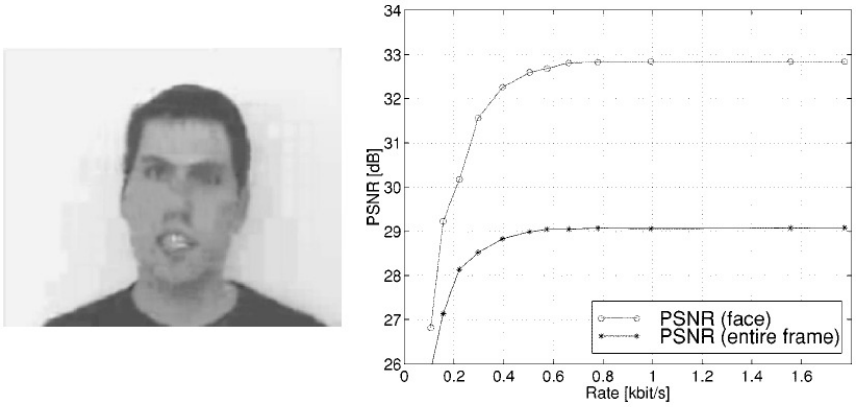


Figure 3.8: PSNR example of compressed sequence. Example of a) a video footage with a moving face on a static background (left image) and b) the PSNR values plotted against a range of kilobits per second for the face alone and the entire image/frame (right image). From Eisert *et al.* (2000) [162].

Image fidelity is concerned with the ability of a distortion process to reproduce an image without any visually visible distortion [186]. For example, when there is not a visible difference between an original image and its compressed version, then the compression method applied is considered to be visually lossless [187]. Fidelity measurements are mainly subjective and can be utilised to determine minimum visual differences (i.e. they determine relative thresholds) (see Section 3.4.2). They effectively identify threshold levels, where the visual changes are small. Objective fidelity metrics exist too, and in great numbers, but they are valid only if they are shown repeatedly to correlate with visual results. These are often very complicated and designed for specific applications [184]. Often, subjective results from fidelity measurements are compared with results from image distortion metrics.

Image fidelity does not always correlate with high image quality. For example, a slightly brighter reproduction (by adjusting the brightness levels) of a dark original will reveal more information and as a result increase the perceived image quality of that image for police tasks (see Figure 3.9). It is unknown if a brighter reproduction of a dark original will improve performance of automated systems. For example, performance of automated face recognition systems does improve when applying illumination normalisation techniques for optimising facial images [188–193]. However, as performance of automated systems is assessed from large datasets without specifying individual scene content properties, it is unknown how these face normalisation illumination techniques influence performance based on specific scene content attributes (e.g. over and/or under exposed scenes).

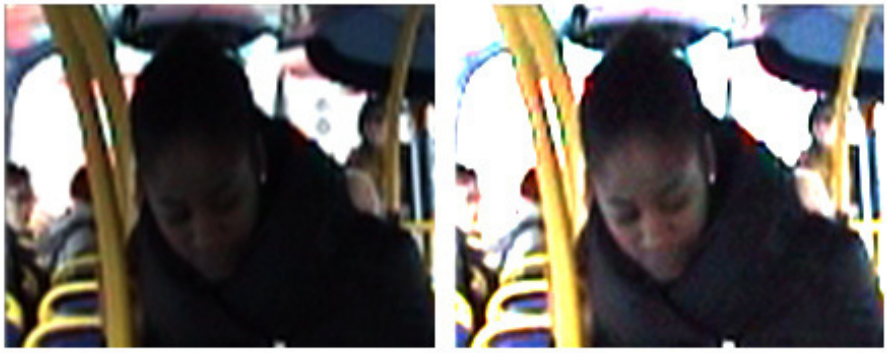


Figure 3.9: Example of a) an original image from a camcorder (left image) and b) a manipulated brighter version of the original (right image). Image b reveals more information.

Fidelity and distortion measures frequently require the presence of the original image, which the distorted (or processed) version is compared against [194]. In image quality measurements the original may, or may not be present, depending on the imaging application. For instance, a reference original image does not exist when assessing the reproduced quality of a camera lens. In this case, the image quality of the lens is evaluated by judging resulted images, either using subjective measures, or image quality metrics (objective measures that are shown to correlate with subjective results) [114, 195–198].

3.3.2 Fidelity, Usefulness and Naturalness

The term image quality, as discussed above, is often considered a general term due to the multiple applications and the broad nature and disciplines relating to imaging. According to Bilissi [20] "observers take into account the purpose, or context for which the image is being used and therefore the same image may be judged differently by different observers, or under different context and conditions". For example, a portrait would be judged differently in the arts context (i.e. image quality in terms of aesthetics) from police applications (i.e. image quality in terms of visibility of useful facial information that could lead to recognition). The FUN model (FUN standing for *Fidelity*, *Usefulness* and *Naturalness*) by Yendrikhovski [18] incorporates the overall quality of an image as a weighted sum of 3 FUN *cognitive dimensions* (see Figure 3.10). The FUN dimensions are under the category of visuo-cognitive attributes (or psychological image attributes) and are only evaluated subjectively [199].

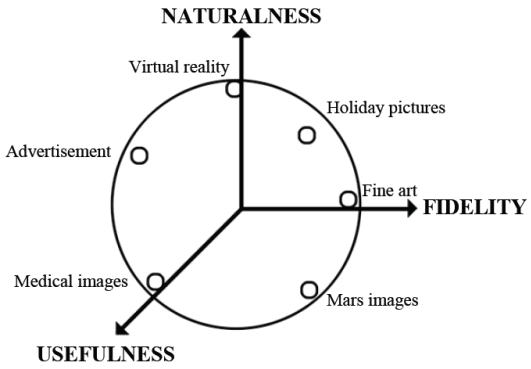


Figure 3.10: The FUN model. From Yendrikhovskij (2002) [18].

Fidelity according to Yendrikhovskij is defined as "the degree of apparent match of the reproduced image with the external reference" [18] and relates to "how accurately we can render an image, without any visible distortion of information loss" [186]. This attribute agrees with the general description of image fidelity in Section 3.3.1.

Usefulness in the FUN model is defined as "the degree of the apparent suitability of the reproduced image to satisfy the corresponding task" [18, 200]. Subjective

image quality for security systems has been described by ITUT P.912 [14] as "the usefulness of the video material to complete a task and not the quality of the video itself". The same definition of image quality is also employed for automated systems [16]. Usefulness is one of the most important attributes for this project because the aim of CCTV imagery is the maintenance of the relevant information for the completion of police tasks (identification/recognition of person, or object, or action).

Naturalness is defined as "the degree of apparent match between the reproduced image and internal reference" [18]. For example, the naturalness of colours to dichromats (people with 2 kinds of cone cells) differs from trichromats (people with 3 kinds of cone cells) [201]. Most people are trichromats.

The use of all the 3 FUN attributes to evaluate overall image quality might not be relevant for a specific application. For example, in surveillance situations, contrast enhancement manipulation techniques might be used in order to enhance visual detail. This could increase the usefulness and as a consequence decrease the naturalness and fidelity. Often, the FUN attributes are evaluated in isolation.

The image quality of security imagery is better fitted under the image usefulness dimension (or attribute). As it has been mentioned in Chapter 2, the human users of security footage (i.e. police staff and specialists) examine in detail the relevant information within the footage. So, image usefulness for CCTV footage is based on the visibility of information that could lead to recognition (e.g. person, action, object). It is not necessarily affected by any artefacts that may disturb the visual image quality (decrease fidelity), as long as these artefacts do not eliminate the relevant information. An analogue of this concept is fingerprint recognition. It has been shown that compression artefacts that are visible in a compressed fingerprint image do not render the image less useful, or of lesser quality than an 'uncompressed' original, as long as the artefacts have not affected important fingerprint ridges that are used in recognition [202].

The visibility of useful information can be eliminated by numerous parameters that influence the initial captured information from the CCTV system and later by storage (or transmission). Some of the most important ones are listed below:

- Object/subject to be recognised in the scene (i.e. clothing, face, vehicle and knife).
- Subject to camera distance (e.g. a close distance may allow facial recognition and a further away distance may allow gait or clothing recognition).
- Angle of camera to the subject (e.g. frontal face view, tilted face view).
- Illumination conditions (i.e. intensity, colour, angle of illumination).
- System performance (i.e. sensor, lens, image processing).
- Recording/transmission (i.e. spatial and temporal compression).

The same term of image usefulness (or utility) is also rightfully employed for automated systems [16, 17] as they complete recognition tasks. However, the usefulness of an imagery for automated recognition systems should be derived based on its performance (hit and miss). This becomes more complicated as recognition algorithms do not work the same as they incorporate different techniques. For example, a certain face recognition algorithm might be designed to be insensitive to illumination changes, whereas such illumination changes might affect severely another face recognition algorithm.

Commonly automated systems are assessed based on correct recognition/detection with ground truth data and with the use of large datasets with minimum scene content classification. For example, a large facial dataset might be categorised to smaller datasets that include illumination variation, differences in ages and more; individual scene content properties are not taken into consideration (see Section 2.1.2). Knowledge of subjective investigations can provide another perspective of testing and/or analysing automated systems. For instance, in Chapter 5 the investigation relates to automated face recognition systems and a testing approach similar to the human fidelity investigations has been implemented (i.e. compressed

versions of scenes are compared against their 'uncompressed' reference).

3.3.3 Basic image attributes

Table 3.1 presents 5 basic image quality attributes that need to be considered when conducting investigations with imagery as they affect the visibility of content information. The basic image attributes can be measured objectively and subjectively (i.e. human investigations). These are *tone*, *colour*, *resolution*, *sharpness* and *noise*. Additionally, Table 3.1 provides the visual description of each basic image attribute together with their associated objective measures. These attributes are effectively related to the performance of imaging systems (e.g. colour reproduction of a sensor) or processes (e.g. sharpening filters) and their derived measures are affected differently by different scene contents. All the basic image attributes are explained briefly in the following paragraphs.

Image Attribute	**Visual Description**	**Objective Measures**
Tone	Macroscopic contrast or reproduction of intensity	Characteristic curve, density differences, transfer function and OECF, contrast, gamma, histogram, dynamic range.
Colour	Differences in lightness, chrominance and hue	Spectral power distribution, CIE tristimulus values, colour appearance values, CIE colour differences.
Resolution	Discrimination of fine detail	Resolving power, imaging cell, limited resolution.
Sharpness	Microscopic contrast or reproduction of edges	Acutance, ESF, PSF, LSF, MTF .
Noise	Random and non-random spurious information	Granularity, noise power spectrum, autocorrelation function, total variance (σ^2_{TOTAL}).

Table 3.1: The 5 basic image quality attributes with their associated visual description and objective measures, adopted from Ford and Triantaphillidou [114, 199].

Tone is concerned with the reproduction of intensities or scene luminance (e.g. reproduction to a display monitor). It can be quantified from observers judgements (e.g. subjective impression of brightness and contrast reproduction) and from objective measures (e.g. luminance differences between original and reproduced images). Tone reproduction is considered the most important image attribute, since

the achromatic visual channel conveys most of the visual information [203]. The objective tone reproduction of an imaging system is described by a transfer function [204] (known as the characteristic curve in analogue imaging systems [205]), in which the output intensity values are plotted as a function of input intensities. From such curves, a measure of global contrast, known as gamma, can be deduced [206]. Tone and contrast (i.e. the difference of intensities between 2 different areas in an image) are inter-related [207]. For example, perceived image contrast depends on the luminance and spatial content structure of an image [208].

Colour is defined as a visual sensation resulting from the interaction of 3 components: the light source, the object itself and the human visual system [209]. There are 3 main perceptual attributes that may describe the perception of colour. These perceptual attributes are hue (i.e. "the human sensation according to which an area appears to be similar to one, or to proportions of two, of the perceived colours red, yellow, green and blue" [210]), brightness (i.e. an area that appears to emit more or less light) and colourfulness (i.e. the perceived area appears to be more, or less chromatic) [211]. Objective colour reproduction is traditionally measured with colorimetry. Colorimetry is based on the theory of trichomatic vision and the spectral sensitivity of the cones of the human visual system (HVS) [212,213]. Often, the colour reproduction is assessed using colour difference models such as CIELAB ΔE^*ab, CIELUV ΔE^*uv, CIEDE2000, iCAM, and iCAM06 [214–216].

Sharpness is defined as the ability of the imaging system to accurately reproduce edges. The subjective impression of sharpness depends on the reproduction of a physical edge and contrast [114]. Image contrast has been defined as the perception of spatial variation [217] such as a text on a uniform background. When the spatial variation is high (e.g. black and white shades) than the image will be perceived sharper from an image of low spatial variation (e.g. mid grey shades). The most common objective method of sharpness measurement is deriving the system's Modulation Transfer Function (MTF) [218]. The MTF describes the contrast reproduction as a function of all available spatial frequencies [219]. MTF measurements

involve the capture of test charts containing sine waves, square waves, slated edges and dead leaves [220–224]. Other methods of measurement involve the system's point spread function (PSF), or the Line Spread Function (LSF), which are both related to the MTF [225]. Sharpness measurements are discussed in more detail in Section 3.3.5.

Resolution is concerned with the reproduction of fine detail and is affected by tone, sharpness, contrast and noise [114, 226]. In analogue imaging, resolution is commonly measured using test charts of line/bar pairs (i.e. known as the resolving power measure). In digital imaging, the system's effective resolution can be deduced from the measured MTF [227]. Digital image resolution, on the other hand, is expressed by the number of pixels in the horizontal and vertical dimensions. For example, a digital sensor of 3000 by 4000 pixels (i.e. in terms of number of photosites), or a displayed image of 600 by 400 pixels. This should not be confused with the 'true' effective resolution of the imaging system, which relates to the smallest image point the system can produce. Another method of measuring system resolution is the PSF (i.e. the response of an imaging system to a narrow point of light).

Image noise is defined as the "unwanted random fluctuation of light intensity" in an image [228]. Noise obscures image detail and it is more visible in spatially uniform areas (i.e. a cloudless blue sky). Noise can be introduced by the imaging system, processes and the signal itself (input intensity). In digital imaging, electronic, photoelectronic and quantisation are common noise sources. In subjective evaluations, noise is defined as graininess (i.e. for analogue systems), or noisiness. Objective measures of noise involve statistical descriptors, such as the standard deviation or the variance, calculated from an imaged uniform field [229]. Furthermore, digital image artefacts are considered 'noise' as they are the products of the system and are not present in the imaged signal [199]. See Section 3.2.2 for more details on digital artefacts.

Sections 3.3.4 and 3.3.5 provide further details on the methods used in the exper-

imental parts of this thesis for colour conversions, sharpness and resolution measurements.

3.3.4 Colour: CIELAB space

The International Commission on Illumination (CIE from its French title, the Commission Internationale de l ′ Eclairage) organisation introduced in 1976 the CIE $L^*a^*b^*$ or CIELAB colour space [230, 231]. CIELAB is a perceptually (nearly) uniform space, which is calculated from the tristimulus values (see Equations 3.1 to 3.3) [230, 232] of a colour and takes account the tristimulus values of the reference white (e.g. illumination). It is a device-independent colour space, meaning that it is not tied to a particular imaging device. Perceptually uniform means that numerical magnitudes correspond to proportional perceptual magnitudes all over the colour space. The CIELAB space has a uniform lightness scale coordinate, L^*, and 2 colour coordinates, a^* and b^* (see Figure 3.11). L^* values range from 100 (white) to 0 (black). Both a^* and b^* coordinates are bounded by the CIE XYZ spectrum locus and represent redness-greenness and yellowness-blueness respectively.

The $L^*a^*b^*$ coordinates are calculated from CIE 1931 X, Y, Z tristimulus values, which are in turn calculated using the CIE 1931 $\overline{x}, \overline{y}, \overline{z}$ colour matching functions. The trichromatic process of the spectral sensitivity of the cones in the human visual system is represented by the colour matching functions CIE 1931 $\overline{x}, \overline{y}, \overline{z}$ [230, 232]. The CIE 1931 X, Y, Z tristimulus values are calculated by the following equations:

$$X = K \int_{340}^{780} R(\lambda) I(\lambda) \overline{x}(\lambda) d\lambda \tag{3.1}$$

$$Y = K \int_{340}^{780} R(\lambda) I(\lambda) \overline{y}(\lambda) d\lambda \tag{3.2}$$

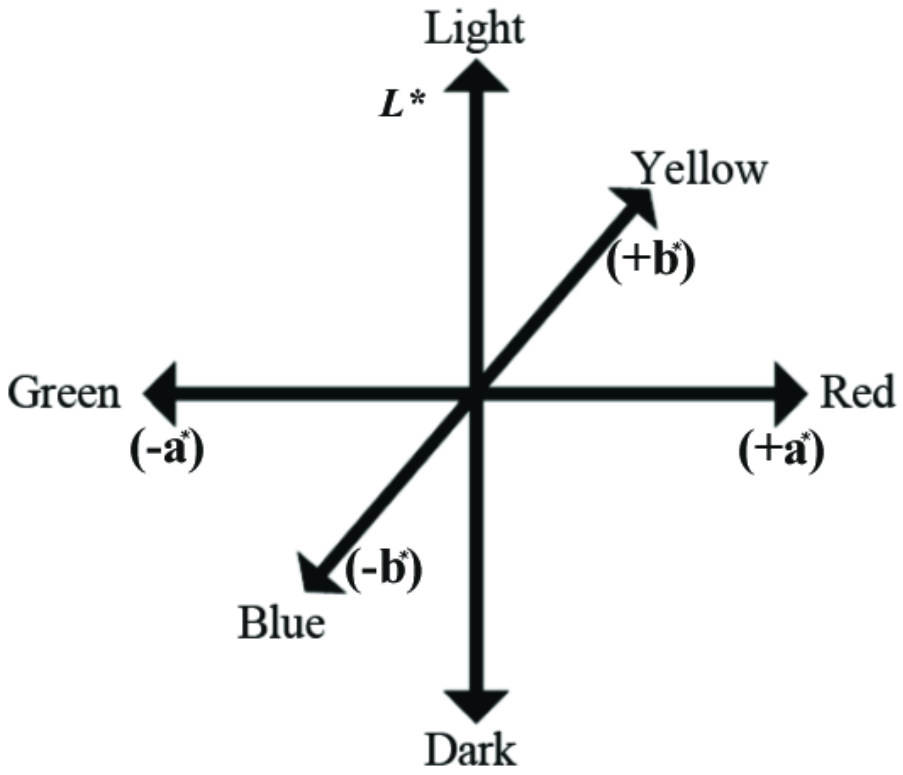


Figure 3.11: The CIELAB colour space. From Fairchild (2005) [233].

$$Z = K \int_{340}^{780} R(\lambda) I(\lambda) \overline{z}(\lambda) d\lambda \tag{3.3}$$

$$K = \frac{1}{\int_{340}^{780} I(\lambda) \overline{y}(\lambda) d\lambda} \tag{3.4}$$

where 340 to 780 nm represent the wavelength range, R is the spectral illuminance, reflectance (or transmittance) of an object, λ is the wavelength of the equivalent monochromatic light, and I is the relative (or absolute) spectral power distribution of the illuminant. Further, the CIE $L^*a^*b^*$ coordinates are calculated by:

$$L^* = 116 \left(\frac{Y}{Y_n} \right)^{\frac{1}{3}} - 16, for \frac{Y}{Y_n} > 0.008856 \tag{3.5}$$

$$a^* = 500 \left[\left(\frac{X}{X_n} \right)^{\frac{1}{3}} - \left(\frac{Y}{Y_n} \right)^{\frac{1}{3}} \right], for \frac{X}{X_n} \& \frac{Y}{Y_n} > 0.008856 \tag{3.6}$$

$$b^* = 200\left[\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - \left(\frac{Z}{Z_n}\right)^{\frac{1}{3}}\right], for \frac{Z}{Z_n} \& \frac{Y}{Y_n} > 0.008856 \tag{3.7}$$

$$L^* = 903.3\left(\frac{Y}{Y_n}\right), for \frac{Y}{Y_n} \leq 0.008856 \tag{3.8}$$

$$a^* = 500\left[7.787\left(\frac{X}{X_n}\right) + \frac{16}{116}\right] - \left[7.787\left(\frac{Y}{Y_n}\right) + \frac{16}{116}\right], for \frac{X}{X_n} \& \frac{Y}{Y_n} \leq 0.008856 \tag{3.9}$$

$$b^* = 200\left[7.787\left(\frac{Y}{Y_n}\right) + \frac{16}{116}\right] - \left[7.787\left(\frac{Z}{Z_n}\right) + \frac{16}{116}\right], for \frac{Z}{Z_n} \& \frac{Y}{Y_n} \leq 0.008856 \tag{3.10}$$

where X, Y and Z are the tristimulus values of the colour and X_n ,Y_n and Z_n are the tristimulus values of the reference white.

In this thesis the L^* value was used to derive an objective measure for characterising the 'lightness properties' of scenes with faces by measuring skin lightness (L^*). *Lightness* is a relative scale that ranges from 0 (black) to 100 (white). Relative scales are useful when the absolute reproduction is impractical. The CIELAB space was chosen over other colour spaces (e.g. RGB non-perceptual and device-dependent colour space) mainly because it provides a perceptual scale for lightness. The alternative would have been to utilise luminance $Y = 0.2126R + 0.7152G + 0.0722B$. More details on the characterisation method are provided in Section 3.5.

3.3.5 Sharpness: MTF evaluation

The modulation transfer function (MTF) is used to assess a system's sharpness; it describes the reproduction of contrast with respect to spatial frequency [219]. It can be applied to colour imaging by treating luminance and colour channels in isolation. The common methods for measuring MTF were briefly mentioned in Section 3.3.3.

In this thesis sharpness measurements of video imaging systems (see Figure 4.1 in Chapter 4) were obtained by employing an adaptation of the edge method, which is specifically for digital images. MTF is more applicable for analogue systems.

The edge method uses a low contrast edge. It is based on the mathematical concept that a 'perfect edge' contains an infinite number of frequencies. Results derived from this method are noisy and often the MTF is overestimated [228]. The work-flow for obtaining the MTF from edges (see Figure 3.12) involves the capture of a 'perfect' edge to produce the Edge Spread Function (ESF, is obtained from scanning, or sampling the edge across); the ESF is differentiated to produce the Line Spread Function (LSF). The MTF is the modulus of the Fourier transform of the LSF (normalised to 1 at $\omega = 0$) [219].

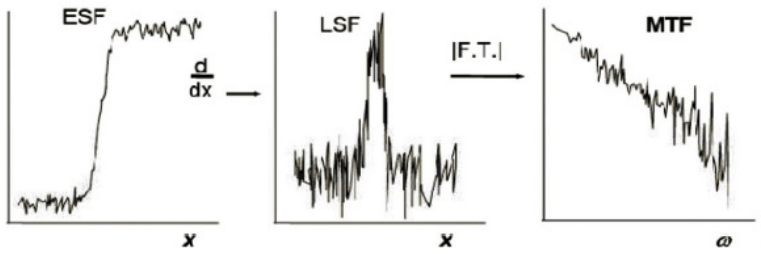


Figure 3.12: The work flow of obtaining the MTF. From Jacobson *et al.* (2000) [219].

In digital systems, an image of an edge needs to be aligned with the pixel array, which makes it hard, if not impossible, to obtain accurate MTF measures. The traditional edge method has been adopted as the slated edge method defined by ISO 12233 [227]. The 'new' slated edge method is based on the old edge method. It derives the so-called Spatial Frequency Response (SFR), which is the equivalent of the MTF, but includes the effect of the slanted edge target in the measured result (i.e. it does not take into account the frequency content of the target). Digital SFRs can be obtained with the use of automated software (P.Burns sfrmat 2.0 Matlab library, or other). SFR measures involve the capture of a low contrast slated edge and greyscale densities (see Figure 3.13). The latter is used to derive the Opto-Electronic Conversion Function, (OECF) of the imaging system, which is in essence the system's transfer function (or tone reproduction function, see Section 3.3.3).

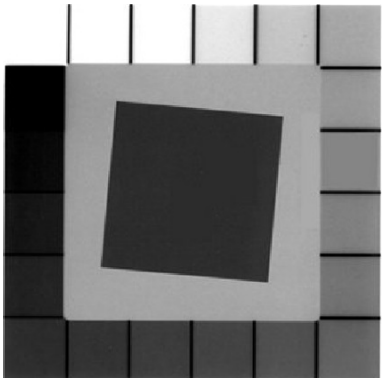


Figure 3.13: A test chart for measuring MTF (using the slanted edge in the middle) and OECF (using the grayscale densities), from Applied Image Inc. (2013) [234]

The OECF is used for the linearisation of the digital data. Digital sensor data, although originally linear, are often processed using non-linear transfer functions for appropriate output/viewing [235]; they are also non-stationary and anisotropic. MTF (and SFR) measurement is strictly applied to linear, stationary and isotropic systems [114]. Nevertheless, providing data linearisation, such measurements are used extensively in the evaluation of digital imaging system performance.

The point at which the SFR drops to 0.1 is a measure of the system's limited resolution (spatial frequency beyond which signal is undetectable) and the point at which the SFR drops to 0.5 is a measure of the system's sharpness (low to medium spatial frequencies are used to quantify sharpness) [236]. Figure 4.1 in Chapter 4 illustrates an example of measured SFRs of a CCTV camera and a DV camcorder.

3.4 Image psychophysics

Subjective image quality investigations involve the use of psychophysics. Psychophysics according to Gescheider [237] is "the scientific study of the relation between stimulus and sensation". In terms of imaging, the stimulus is the still image (or video footage) and the sensation is expressed as the observers' response to a task or question. Psychometric scaling, or psychological rulers are employed to quantify the observers' responses. Psychophysics, when conducted with detailed

planning, produce accurate quantitative results from qualitative responses [169]. The observers' responses are affected by many variables such as image attributes, the quality criteria of the observer, and viewing conditions [238, 239]. Engeldrum [238] has identified some basic steps to guide imaging product developers for assessing image quality. These are:

- Selection and preparation of samples (stimuli).
- Selection of observers and determination of observers' task or question.
- Presentation of the sample to the observers and collection of their responses.
- Analysis of observers' responses and generation of scale values.

The following sections describe the rulers (measuring scales) for collecting subjective responses (Section 3.4.1) from observers and the common investigative methods that are used in psychophysics (Section 3.4.2). Further, Section 3.4.3 presents the psychometric curve data analysis method employed in Chapter 4. Section 3.4.4 focuses on the methods that can be used to assess image usefulness for police tasks.

In terms of automated systems perhaps similar or adapted methodologies from human investigations can be utilised when testing performance of systems where the observers' psychophysical response is replaced with a numerical output.

3.4.1 Measurement scales

About 4 common types of measurement scales or psychometric scales exist [240]. These are nominal, ordinal, interval and ratio (see Figure 3.14). The scale types range in mathematical strength and complexity from nominal to ratio. Also, observers' skills and amount of data complexity increase with the increase of the scale types from nominal to ratio [171]. For example, a ratio scale will have the properties of all the previous scales (i.e. nominal, ordinal and interval), which could be derived by using different analysis and data reduction methods [238].

Nominal scales allocate labels, or names to the content of a sample. This method

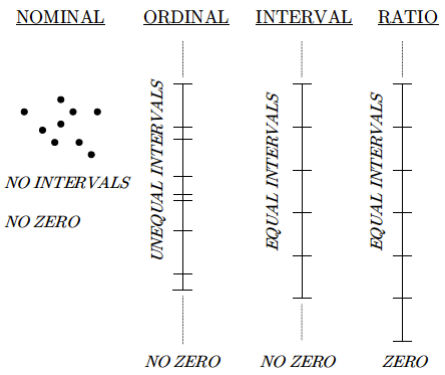


Figure 3.14: Measurement scales. From Triantaphillidou (2011) [199].

of scaling is useful for classifying images based on labels (or names). For example, it could be used to classify images based on categories such as portraiture, architectural, textural, natural and so on [241].

Ordinal scales use numbers or labels in order to rank images according to an attribute. For example, images can be ranked based on the sharpness attribute (e.g. ranking of images from low sharpness to high sharpness). Ordinal scales have the property of 'greater than' or 'less than' depending on the direction of the ranking. The main limitation of the ordinal scale method is that perceptual magnitudes do not correspond to the numbers in the scales and thus the numbers have only the property of 'greater than' or 'less than' [241].

Interval scales use equally spaced intervals corresponding to equal perceptual magnitudes. For example, the equal differences in scale values represent equal perceptual differences of an image attribute or overall image quality. The interval scaling method answers the 'how close' question. Not all image attributes have a fixed zero point, so interval scales provide relative values (i.e. they are 'floating scales') [171, 241].

Ratio scales are interval scales, but with a fixed zero origin. The zero origin causes some experimental difficulties, as some attributes could not start with a zero origin. For example, it is difficult to define zero quality [199, 241].

3.4.2 Psychophysical methods

Various psychophysical methods are used in subjective investigations to derive psychometric scales [242]. The choice of the method depends mainly on the intention of the investigation. For example, fidelity investigations involve threshold evaluations and image quality investigations involve supra-threshold evaluations.

Supra-threshold methods [242,243] are used to derive psychometric scales relating to an image attribute (i.e. the 'ness'), or overall image quality. The 'ness' refers to the ranges of an image attribute such as sharpness or colorfulness. In this investigation the 'ness' represents amounts of compression and/or information content of scenes (i.e. facial information). The most common supra-threshold methods are: 1) rank order (derives mainly ordinal scales): the observers are asked to rank image samples based on the order of an attribute such as text darkness, 2) paired comparison (derives ordinal scales that can be transformed into interval and ratio scales): the observers select 1 image from a paired sample that has more of the 'ness', or it is the preferred from the pair, and 3) category methods (best for deriving ordinal scales): observers see 1 sample at a time and are asked to place it to a named or numbered category.

Threshold investigations [242, 243] are used to identify the just detectable (or just noticeable) amount of an attribute (i.e. value of 'ness' that it is just visible/detectable). Just noticeable difference (JND) investigations are used to identify attribute differences (minimum value of 'ness' that it is seeing as different from a standard) [243]. Some common threshold investigation are: 1) Method of limits (derives results for both threshold and JNDs): the observer is presented with a sample that the 'ness' is imperceptible/clearly perceptible and the 'ness' amount is increasing/decreasing until the observer can detect/cannot detect the 'ness' attribute, and 2) Method of adjustment. This is very similar to the method of limits. The difference is that the observer is adjusting manually the 'ness' by using control methods such as turning a knob or moving a slider.

In the human investigation in Chapter 4, the threshold method of constant stimuli

is employed. This method is used to derive results for both threshold and JND difference. A constant set of samples is used that remain fixed throughout the experiment. The set samples cover a range from low to high level of 'ness' (e.g. levels of compressed amounts). The set samples are selected so that the low 'ness' samples are never selected and high 'ness' samples are always selected by the observers. This method collects data that is analysed with a psychometric curve (see Section 3.4.3). In threshold investigations the reference is not provided and observers answer with a *yes* when they can see the 'ness' [243, 244]. Whereas in JND the reference is provided, the observers compare each sample with the reference and indicate if the sample conveys more of the 'ness' than the reference.

3.4.3 The psychometric curve

The psychometric curve describes the proportion of observers' *yes* responses to different 'ness' levels (e.g. different levels of compression or information content) [245]. Observers are presented with a stimulus and asked to respond with a *yes* or *no* if they see a 'ness', or if they see a difference between stimuli. This process is repeated for a range of 'ness' levels. A *yes* response scores 1 and a *no* response scores 0. The proportion of *yes* responses is the sum of all the responses divided by the number of observers. Statistically, the curve is called cumulative density function and the observers' response of *yes* or *no* is a random variable (it is conceptually identical to the supra-threshold pair comparison method) [243]. Psychometric curves are used to determine thresholds, JNDs, i.e. fidelity measures. A typical psychometric curve is shown in Figure 3.15.

The absolute threshold (i.e. smallest amount of the 'ness' required to see the 'ness') is at the point where 50% of the observers responses are *yes*. The just noticeable difference (JND) represents the stimulus change required to produce a JND of a 'ness'. It is typically defined at the point where 75% of the responses are *yes*, but this percentage of *yes* replies is not universal [243]. In some applications 0.60 or 0.66 has been taken as the JND. The point of subjective equality (PSE) is the 0.5

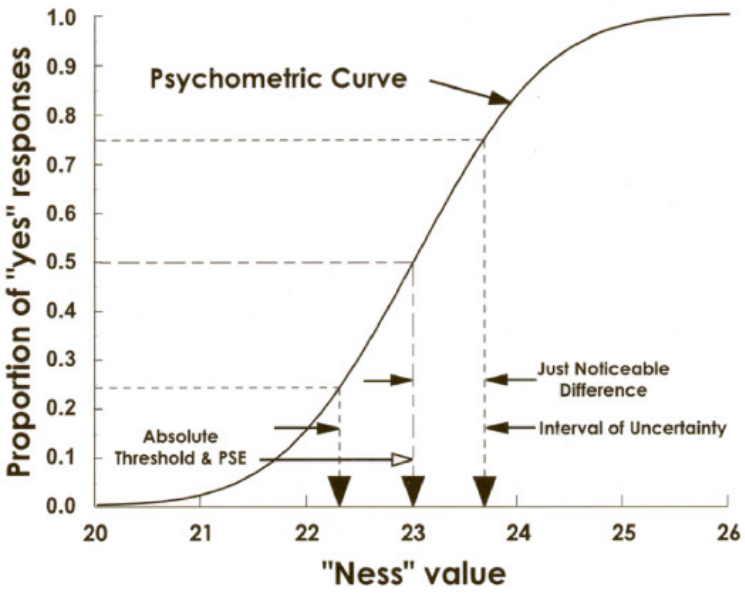


Figure 3.15: The psychometric curve. The x- axis shows the range of the 'ness' levels and the y-axis the proportion (or probability) of observers *yes* responses. From Engeldrum (2000) [243].

point on the y-axis and relates to when observers find 2 stimuli equal in statistical terms. The range between 0.25 and 0.75 points of *yes* responses is referred to as the interval of uncertainty, where responses could go either way (*yes* or *no*).

3.4.4 Image psychophysics for recognition tasks

In the evaluation of perceived image usefulness, for recognition tasks, a reference may or may not be included. Figure 3.16 provides an example were the observer is presented with a single facial image and asked to judge the image usefulness using a category scale from 1 (the lowest quality) to 5 (the highest quality) [44].

The ITU-R BT.500-11 provides guidance in relation to assessments of video based imagery [194]. When image fidelity is assessed, the ITU recommends that the reference video is provided and runs simultaneously on a single monitor, along with the reduced quality version. This arrangement helps the observers to make a direct judgement of what they see, and does not rely on memory. A similar methodology to the fidelity assessment is used for the assessment of image usefulness in Chapter 4. The observers had to answer with a *yes* or *no* to the question "Is the reproduced

Figure 3.16: Single image perceived image quality. An example of single image perceived image quality (i.e. more specifically image usefulness). From Adler *et al.* (2006) [44].

image as useful in terms of facial information as the reference?" (see Figure 3.17).

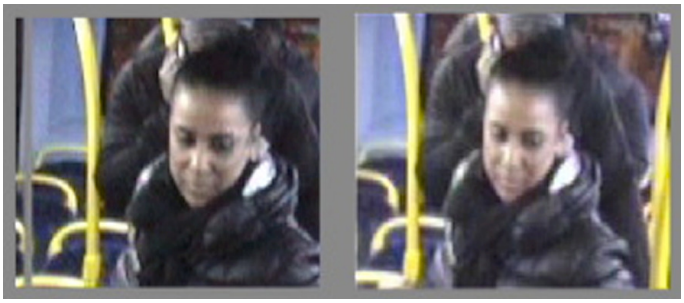


Figure 3.17: Image quality when the reference is provided. The information within the impaired version (image on the right) is judged against the facial information from the reference (image on the left).

The ITU-T P.912 Recommendation provides subjective assessment methods for 'target' recognition video (TRV) based tasks [14]. TRV methods can be used for a variety of purposes, such as human recognition, licence plate recognition, remote monitoring and decision making. A couple of main methods are recommended, the *multiple choice* method and the *single answer* method. The stimuli used (i.e. the samples) in the evaluations should reflect operationally the conditions of the collected video material (e.g. transmission service under test) and cover all possible scenarios for the particular application. Also, the methods "assess the ability of the viewer to recognise the appropriate information in the video, regardless of the viewer's perceived quality of the viewing experience" [14].

- Multiple choice method. In this method the observer is presented with a particular target category (human, object or/and alphanumeric, action, scars, tattoos and so on), which need to be recognised in a video. After the video presentation the observer chooses the label, which corresponds closest to what they see in the video (see Figure 3.18). The answers are either correct or incorrect. The derived data are analysed by examining the stability of the answers within and between subject performances.

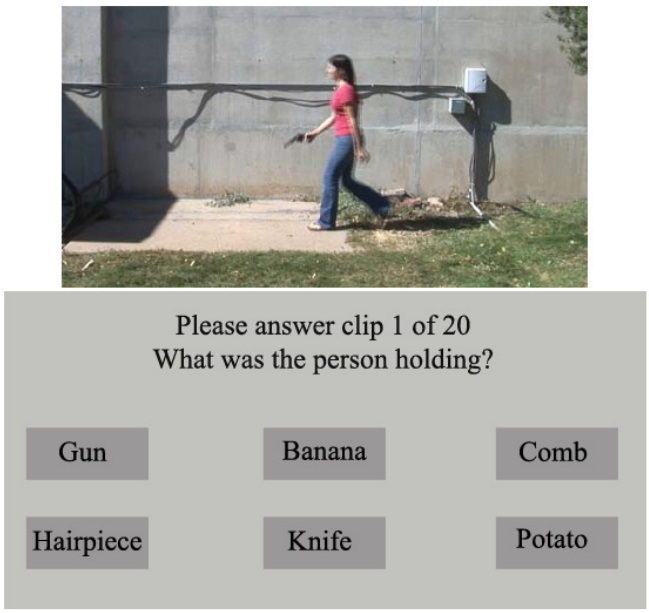


Figure 3.18: Example of display of the multiple choice method. From ITU-T P.912 (2008) [14].

- Single answer method. In this method, the observer is asked to type the letter(s) or number(s) present in a specific area of the video (e.g. numbers of a number plate) (see Figure 3.19). Additionally, under the same method the observer can answer with a *yes* or *no* to the question, if a certain target was present in the video. The answers are either correct or incorrect. The derived data are statistically analysed by determining the observer's performance above the level of chance for answering correctly.

Finally, the face image matching abilities of observers can be derived by the use of neuropsychological tools such as the Benton Face Recognition Test (BFRT) [246]. The test requires the identification of a target face from a set of faces. Figure 3.20

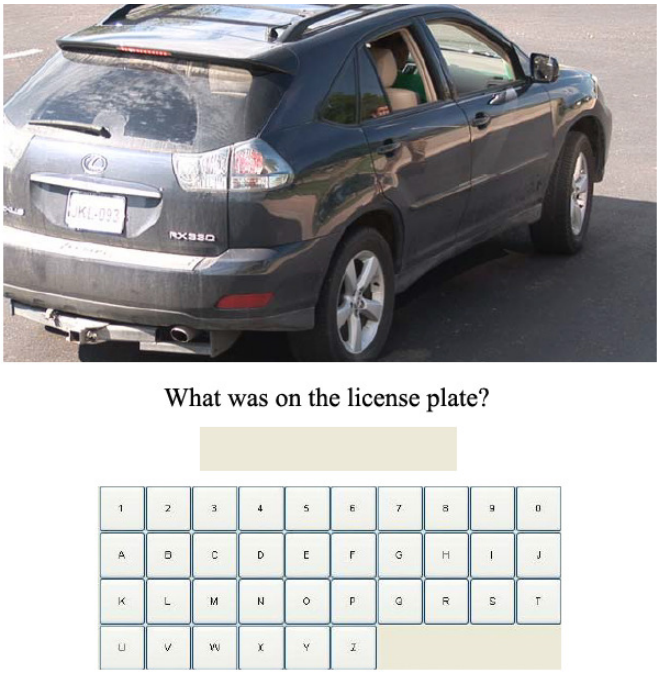


Figure 3.19: Example of display of the single answer method. From ITU-T P.912 (2008) [14].

illustrates an example. The observers are required to select 3 photos that depict the face photo number 7 (the correct answer is 2, 5 and 6) [246].

3.4.5 Other factors influencing measurements

Apart from the selection of a suitable method to collect psychophysical data, there are a number of other factors that might affect the reliability of the collected data. These include: type of observers, number of observers, the observer task instruction, choice of test sample, viewing conditions, environment (temperature and noise), the presence of the experimenter in the room or not, and observer fatigue mainly due to the length of the experiment.

Normally observers fall into 2 types: lay or expert. Expert observers have experience in judging, or evaluating images and lay observers do not. To date, there is not much research about whether people with experience in forensic image matching tasks (i.e. police officers) are any better from untrained civilians. A couple of studies have been located that provide a mixture of evidence for matching unknown faces from images. In one study, police officers had the same low accuracy as civilian participants in

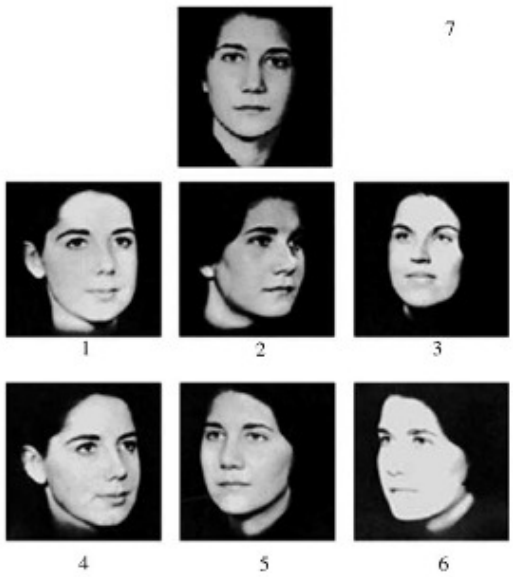


Figure 3.20: An example of the FFRT test. From Duchaine *et al.* (2003) [246].

identifying an unknown target face from CCTV footage [9]. In another study, expert facial image comparers had a higher accuracy in matching of faces, between photos and CCTV footage, than civilians [247]. It should be noted that there are many differences in the respective methodologies between these 2 studies, which makes it difficult to reconcile the results (e.g. people with different experiences, different image samples and different experimental methodologies).

Having a larger number of observers decreases the errors of the collected data and thus increases the precision of the data. The recommended range of observers is between 10 to 30 for a typical psychophysical investigation [238]. A more precise observer number estimate can be derived from establishing the desired scale precision. The ITU-T P.910 recommends a range of observers between 4 (as the minimum) to 40 (as the maximum) [14], depending on the application of the investigation. For example, small group of experts (4 - 8) can provide indicative results [14].

Observers' instructions should be preferably written and include information on what they are judging and how to respond. The same instructions should be given to all the observers in order to retain consistency. Common practice involves the training of the observers with a small number of representative scenarios from the actual test. The results from the training scenarios should not be included in the

analysis.

The choice of image sample (stimulus) is determined by the objective of the psychophysical investigation. After the selection of the scaling method the experimenter would have to make decisions on the range of distribution of the 'nesses' and the image content of the sample. If a particular application were under test, then the selected sample would need to be representative of that application.

Standard and controlled viewing conditions are vital to enable reproducible psychophysical data. For example, the monitor would need to be calibrated and the illumination and surrounding condition in the experimental room would need to be controlled. As well as, viewing distance and the screening of observers for normal vision (e.g. corrected vision by wearing glasses) are some other factors that need to be taken into consideration.

3.5 Scene dependency and classification

Results from subjective investigations and performance of automated systems are often shown to vary with scene content. This is known as scene dependency. All the background research in Chapter 2 affirms the scene content dependency essence of police tasks undertaken by human operatives (see Section 2.1.1) and automated systems (see Sections 2.1.2 and 2.1.3). Scene content properties/characteristics do not only affect police tasks but also compression algorithms themselves. In Section 3.2.3, background research proves that compression performance is also influenced by scene content properties. For example, the ringing compression artefact is more obvious around very steep edges and often in natural images is not visible [248]. In fact, any image transformation/process is affected by scene content. Another example, is the basic image attributes described in Section 3.3.3.

3.5.1 Scene characterisation and classification

Scene dependency can be overcome with the use of scene characterisation and later classification (characterised scenes are placed in groups) methods (see Section 3.5.1). Both objective and subjective methods (use of visual/empirical inspection) can be employed for characterisation purposes. The objective methods recommended by the ISO/IEC 29794-5:2010 could be considered more appropriate for the characterisation of CCTV footage. For example, the ISO/IEC 29794-5:2010 describes scene properties/characteristics that will have an effect on the visibility of information within an imagery such as brightness, exposure, camera to subject distance and so on. The methods described in ISO/IEC 29794-5:2010 could be seen to be applicable for any recognition task (i.e. object, action), not just for faces.

The following characterisation techniques have been used, in the experimental part of this thesis, for classifying scenes with facial (for investigations in Chapters 4 and 5) and human silhouettes information (for investigation in Chapter 6). The techniques extract information relating to local (i.e. just on subject to be detected/analysed such as a face or human silhouette) and global (i.e. on the entire scene) scene properties.

I Camera to subject distance. This is a local face characterisation technique and is derived objectively, by measuring manually the inter-pupillary distance in pixels.

II Scene lightness. This is a local face characterisation technique and is derived objectively from measuring skin lightness CIELab L^*. The skin lightness may be affected by both the scene illumination and the colour of the person's skin. Lightness levels ranged from 0 (no lightness black) to 100 (maximum lightness white). An average of 4 areas on the face is used. The areas are the forehead, the right cheek, the left cheek and the jaw. In case of facial hair the jaw area is not measured.

III Angle of face to camera plane. This is a local face characterisation technique

and is derived subjectively by visual inspection. Figure 3.21 illustrates examples of face angles. Images that include most of both cheeks (between -20 and + 20 degrees on the horizontal axes) and the very top of the head is not visible (between 0 and +10 degrees on the vertical axes) are classified as frontal. Images that include most of both cheeks (between -20 and + 20 degrees on the horizontal axes) and the very top of the head is visible (e.g. +20 degrees and above on the vertical axes) are classified as tilted.

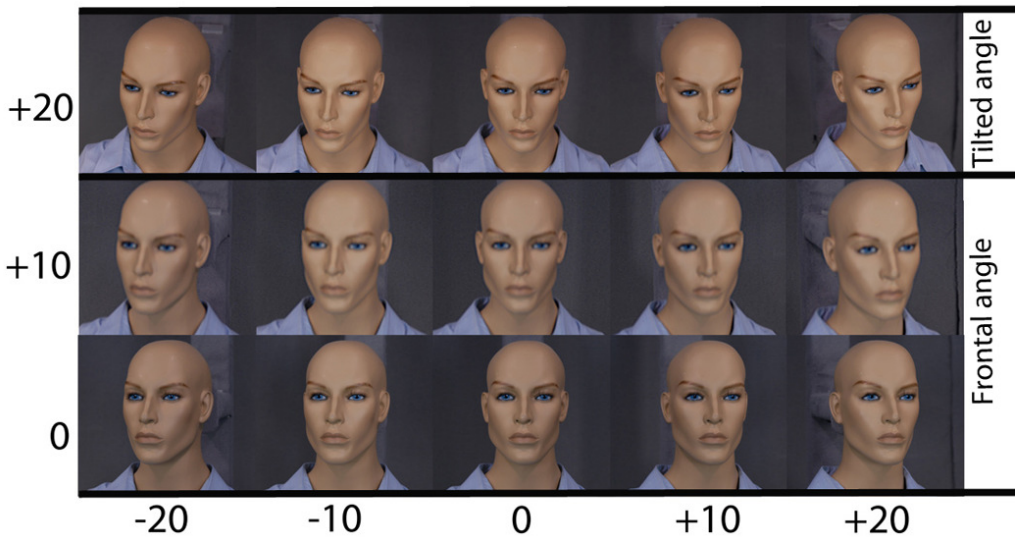


Figure 3.21: Partial groups of facial angles in degrees.

IV Scene contrast. This characterisation technique is applicable for scenes that include human silhouettes (e.g. as the sterile zone scenario in Figure 2.9) and is derived objectively. The subjects (human silhouettes), in the sterile zone scenario wear only 2 types of clothing, white or green. The head of the subjects is excluded from the measurements in order to avoid complications with the derived measures. The method involves a calculation of a ratio of dark to light area between foreground and background. The derived values from the contrast ratio range from 0 to +1 (see Eq. 3.11).

$$CR = \frac{L_{min}}{L_{max}} \tag{3.11}$$

Where L_{max} and L_{min} are the maximum and minimum lightness (L^*) values respectively. The lightness values were derived by measuring lightness in specific

areas in the scene using the CIELAB colour space.

V Scene busyness. This is a global scene characterisation technique and is derived objectively, by measuring the global spatial and temporal properties of the scenes. An objective measure using ITU specifications, is implemented [249]. The spatial information is extracted by using the standard deviation of Sobel filtered frames; the maximum value represents the spatial information for the scene (see Eq. 3.12). The temporal information is obtained using the standard deviation of the frame differences; the maximum value represents the temporal information for the scene (see Eq. 3.13 and 3.14).

$$SI = max_{time}\{std_{space}[sobel(F_n)]\} \tag{3.12}$$

where SI stands for spatial information, max_{time} is the maximum value among the standard deviation (std_{space}) of Sobel filtered frames ($sobel(F_n)$).

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{3.13}$$

$$TI = max_{time}\{std_{space}[M_n(i,j)]\} \tag{3.14}$$

where ($M_n(i,j)$) provides the frame differences among a pair of frames ($F_n(i,j) - F_{n-1}(i,j)$). Where TI stands for temporal information, max_{time} is the maximum value among the standard deviation (std_{space}) of frame differences ($M_n(i,j)$).

3.6 Discussion

The experimental part of this thesis (Chapters 4, 5 and 6) investigates performance evaluation of automated systems and human operatives with aspects of image quality, compression and CCTV imagery. Resources from different academic scientific disciplines (relating to image/video processing and compression, psychophysics, im-

age quality, use of police imagery, human face recognition studies and automated algorithms) have been combined in Chapters 2 and 3 in order to enable the adaptation of appropriate evaluation performance methodologies. For example, any image transformation/process is affected by scene content properties. This has been taken into consideration by including a variety of scene content properties for each investigation. Additionally, as the performance of police tasks is influenced by scene content properties and the different techniques incorporated by the visual systems (more for automated systems), perhaps the use of the term *image acceptance* is more appropriate than image quality and/or image usefulness. The term *Image acceptance* specifies the appropriateness of the scene content property to complete the task for the specified visual system (i.e. algorithm, human visual system).

Image quality investigations aim to describe/measure image attributes (e.g. usefulness, naturalness, sharpness and general image quality) by utilising either objective (e.g. measures of distortion) and/or subjective (e.g. psychophysics) means. The image usefulness attribute for both human and automated recognition systems can be achieved by measuring performance from ground truth data (correct detection or no detection or false detection). For human investigations, experience derived from completing recognition tasks can also be utilised (i.e. use of expert-observers) to further understand behaviours or appropriateness of image quality (see Section 3.4.5). This is not the case for automated systems and their behaviour can only be understood based on numerical performance measurements.

The following 3 chapters are concerned with the experimental part of the thesis. The next Chapter 4 relates to human face recognition, Chapter 5 to automated face recognition, and Chapter 6 to human detection as part of an analytics system.

CHAPTER 4

Case study 1: Identification of acceptable bitrates for human face recognition from CCTV imagery

Human face recognition from images or video footage requires a certain level of recorded image quality. This chapter derives acceptable bitrates (relating to levels of compression and consequently quality) of footage with human faces, using an industry implementation of the standard H.264/MPEG-4 AVC and the CCTV recording systems on London buses. The London bus application is utilised as a case study for setting up a methodology and implementing suitable data analysis for face recognition from recorded footage which has been degraded by compression.

4.1 Introduction

There are many surveillance applications where the relatively accurate recording of facial information is possible, such as in trains, buses, underground, transport stations and open streets. When a person is further away from a camera then less facial information will be visible and other means of person recognition can be applied,

such as gait analysis [250]. In this investigation, the London bus application was selected as a case study. London public buses make use of CCTV systems to prevent crime, recognise offenders/actions and for insurance purposes [251, 252].

This chapter proposes a reproducible methodology and tools to derive acceptable bitrates of scenes with human faces using an industry standard implementation of H.264/MPEG-4 AVC, and the CCTV recording systems on London buses. The majority of CCTV recorders on buses use a proprietary format based on the H.264/MPEG-4 AVC video coding standard, exploiting both spatial and temporal redundancy. Low bitrates are favoured in the CCTV industry for saving storage and transmission bandwidth, but they compromise the *image usefulness* of the recorded imagery. In this context, usefulness is determined by the presence of enough facial information remaining in the compressed image to allow a specialist to recognise a person. The investigation includes 4 steps: 1) development of a video dataset representative of typical CCTV bus scenarios, 2) selection and grouping of video scenes based on local (facial) and global (entire scene) content properties, 3) psychophysical investigations to identify the key scenes, which are most affected by compression, using an industry implementation of H.264/MPEG-4 AVC, and 4) testing of CCTV recording systems on buses with the key scenes and further psychophysical investigations.

Scenes of 20 second duration were grouped using 4 scene classification techniques from Section 3.5.1. These are: scene lightness, camera to subject distance, angle of the face to the camera plane and level of busyness (based on spatial and temporal information). A couple of psychophysical investigations were conducted with the help of experts from the Metropolitan Police Service (MPS) and bus analysts. The first was used to identify the key scenes (i.e. scenes affected most by compression), from an initial selected set of 25 scenes, using an implementation of H.264/MPEG-4 AVC. The second was used to identify acceptable bitrates of the pre-selected key scenes (resulting in a set of 6 scenes), using five of the most commonly used CCTV recording systems on London buses. The former psychophysical investigation

acted as a filter in order to reduce the viewing experimental time in the latter psychophysical investigation. In both investigations, the expert observers had to answer with a *yes* or *no* to the question "Is the compressed version of the scene as useful as the reference original in terms of facial information?".

The findings aim to contribute to optimising the conditions around facial recognition tasks undertaken by specialists, by tuning the compression to a just acceptable level. The rest of the chapter is organised as follows. Section 4.2 contains a description of the experimental methodology. Data analysis of the results and discussion of the 2 psychophysical investigations are provided in Sections 4.3 , 4.4, and 4.5. In Section 4.6 conclusions are drawn.

4.2 Methodology

The acceptable bitrates were derived by carrying out 4 steps: 1) development of a representative video dataset, 2) selection, classification and grouping of video scenes, 3) identification of key scenes using an industry standard implementation of H.264/MPEG-4 AVC, and 4) testing of five CCTV recording systems using the identified key scenes.

4.2.1 Development of a representative video dataset

A sunny day presents challenges in terms of illumination for recording activities on buses. When the sun illuminates one side of the bus, some areas in a scene are over-exposed, while others under-exposed. As the bus moves, the windows allow illumination from different directions, causing the areas of over and under exposure to vary rapidly. In contrast an overcast day will produce diffuse light and uniform illumination and probably correctly exposed scenes, which might not be challenging enough for testing compression. When natural light is low, the bus lighting is the main source of illumination. It was observed that it produces a more predictable and rather uniform illumination. During this time other sources

of lighting, such as street lamps and lights from other vehicles, might influence the properties of the bus lighting. Yet, not to an important degree, as the main bus lighting dominates the scene. The following conditions were used during data collection (footage recording).

- Camera system. A consumer quality mini digital video (DV) camcorder (Sony DCR-HC37E with available focal length distance between 1.9-76mm and focal ratio between f/1.8-f/4.1) was used for the filming of all scenes. Automated settings for exposure, white balance and focus were chosen to replicate what happens with actual CCTV camera capture. The automated settings will have an impact on the produced footage (over or under exposed scenes, out of focus scenes, incorrect white balance) but the aim is to replicate reality and create a representative dataset for the bus application. For example, at lower illumination levels that the bus illumination provides compared to daylight levels, the f/stop will become smaller (i.e. aperture widens), resulting to smaller depths of field. This again will decrease the influence of light sources outside the bus. About 10 camcorders were set up according to Transport for London (TfL) recommendations, i.e. the camera views (see an example in Figure 4.2).

  Although the camcorder has been set to automatic exposure to mimic the cameras installed in the buses, which are also set to automatic exposure, it is not known (or quantifiable) how the automatic camcorder camera settings differ from the bus camera settings, since bus cameras are neither standardised nor characterised prior to installation. White balance differences between camcorder and bus cameras are not expected to have much impact in the results, since white balance is not an essential element in human face recognition [253]. A discrepancy in auto-focus could be possible, but out of focus facts are not part of the experimental parameters in respect to face recognition after vs before compression.

In addition, the camcorders at close distance to the subjects (e.g. doorway view, staircase view) were fitted with wide-angle conversion lenses (SONY VCL-HGA07B 0.7x) to provide a wider-angle of view and compensate for the restricted camera to subject distance area. There was no difference on the obtained SFR measures (see Figure 4.1) between the camcorders with and without the fitted wide-angle conversion lenses.

- Illumination conditions. Sunny day (during day time) and bus illumination together with some exterior illumination e.g. from shops, street lamps and from other vehicles (during night time).

- Participants. 26 actors from various ethnicities, ages and gender acted as the bus passengers, according to given scenarios.

The DV consumer camcorder was chosen for the recording of the bus dataset over a CCTV camera for various reasons, including accessibility, quality and cost. For example, expensive, specialised equipment is required in order to record the output of a CCTV camera in an 'uncompressed' format. Also, there are numerous companies that provide CCTV systems to London buses. These have large variations in quality, which have not been studied and quantified. This also relates to the camera and lens properties of these systems. Often, the camera specifications are not provided as these systems can be bought on-line from other countries (e.g. China) by the companies that provide CCTV systems to the end users.

A typical sample CCTV camera was provided by a supplier in order to examine and compare its wide-angle properties (and SFR measures) to the DV camcorder with and without the wide-angle lens converter. The sample CCTV camera still had slightly wider-angle properties than the DV camcorder with the wide-angle lens converter. Wide-angle lenses are used in CCTV applications in order to cover wider scene content information and they do distort image content. In particular when the subject of interest is not in the middle of the lens. How much a very wide-angle lens distorts content information or affects the correct completion of police tasks is an investigation on its own merits. TfL was planning to change the specifications

for the use of wide-angle lenses and did not wish to include very wide-angle lenses in the investigation. The difference between the DV camcorder with and without the wide-angle conversion lens was noticeable but did not appear to distort the content information as only a 0.7x factor was utilised. Moreover, a very wide-angle lens will make subjects in the scene to appear at a greater distance than a less wide-angle lens. This does not matter in this investigation as scenes are evaluated based on their scene content properties and these have been characterised and grouped.

Figure 4.1 provides a comparison of the Spatial Frequency Response (SFR) [227], of a typical sample CCTV camera used on buses with that of the DV camcorder used for the collection of the dataset. The SFR measure was chosen to assess differences between the aforesaid systems as it provides a measure on the reproduction of fine detail (resolution) and sharpness/contrast (see Sections 3.3.3 and 3.3.5). These 2 contribute the most to the capture of useful facial information as resolution measurements alone are affected by tone, sharpness, contrast and noise. The SFR allows the incorporation of many variables/factors in a simple measure.

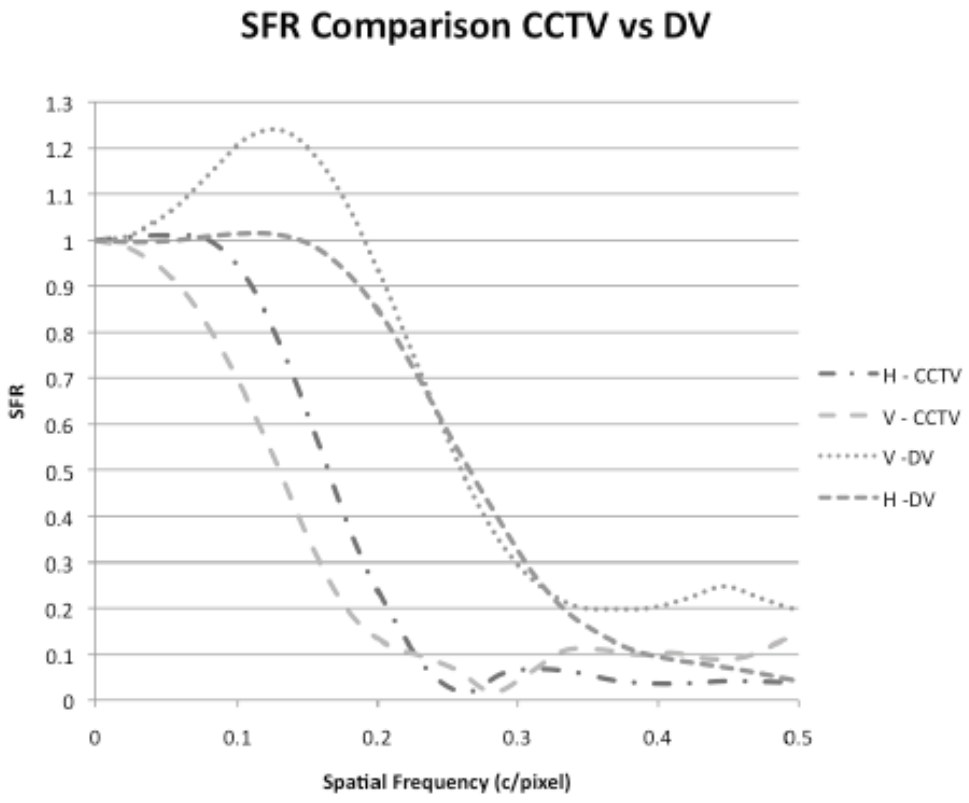


Figure 4.1: Example of a SFR measure. Horizontal (H) and Vertical (V) SFR of a CCTV camera and a DV camcorder.

The DV camcorder SFR indicates image sharpening in the vertical camera orientation, in the low and mid frequencies. Further, the camcorder has a much greater optical resolution (i.e. the SFR falls to 0.1 at nearly 4 pixel-1) and produces sharper images (i.e. 0.5 SFR corresponds to approximately 2.7 pixel-1) than the CCTV (i.e. optical resolution limit at less than 3 pixel-1 and 0.5 SFR at less than 2 pixel-1).

The consumer DV camcorder is shown to have produced overall higher quality output than the CCTV system. One option to compensate for this difference is to apply a frequency filter (i.e. a blurring filter), aiming to visually match the frequency response of the DV recorder to that of the CCTV camera [254]. In this case, this option was omitted, since the current rapid development of CCTV system technology will result in CCTV systems producing comparable image quality to that of consumer video systems. The focus of the work was put on setting up an experimental paradigm in the investigations and implementing a suitable analysis of results.

The footage dataset was recorded in a DV format, at 25 megabits per second (Mbits/s), 4:2:0 chroma subsampling, at full D1 PAL resolution (720 x 576) and with interlaced scanning at 25 frames per second (fps).

Figure 4.2: Example camera views of the CASTBUS 2012 dataset.

The developed video bus dataset is called CASTBUS 2012 and it can be obtained from the Home Office Centre for Applied Science and Technology (CAST) in UK [255], to assist those wishing to investigate solutions in relation to the bus video recordings.

4.2.2 Selection, Classification and Grouping of Video Scenes

In this investigation, various scenes were selected from the CASTBUS 2012 dataset and were further compressed using the MPEG-2 coding standard at approximately 25Mbits/s (4:2:0 chroma subsampling). This compression has enabled the five suppliers of bus CCTV recording systems to have the key scenes on a DVD. The suppliers were asked to play out (with a DVD player) the key scenes into their recording system according to a pre-defined number of bitrates and to return their recordings for use in the experimental testing.

The main difference between DV and MPEG-2 compression is in the temporal domain; otherwise both encoders are based on the DCT transform [256, 257] (i.e. MPEG-2 exploits both spatial and temporal redundancies whereas DV exploits only spatial redundancy). An initial experiment, involving only the experimenter, was conducted to appreciate empirically the visible differences between the 2 encoders. The experiment involved careful observation of a number of compressed scenes, with various scene properties. No visible differences were observed between the compressed scenes. Figure 4.3 illustrates an example comparison using both encoders. The compression bitrates used in the CCTV industry are typically much lower than 25 Mbits/s. Thus, the additional compression of the reference using the MPEG-2 encoder should not affect the results.

Due to the miscommunication between transmission and recording, it was observed that CCTV recording systems sometimes recorded the 2 fields as 1 frame causing the interlace effect (see Figure 3.2). In order to avoid this effect in the compressed scenes, one of the fields (the odd line numbers) was removed and the even number

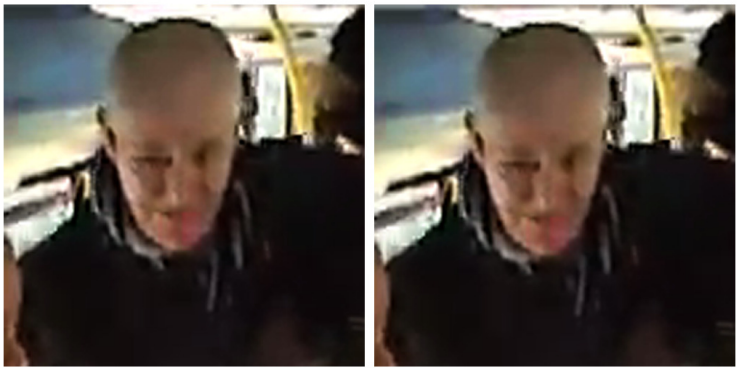


Figure 4.3: Comparison between MPEG-2 and DV encoders. Comparison of 2 images compressed at 700kbps, with MPEG-2 (right) and DV (left).

of fields were interpolated using the $Apple^{TM}$ Final Cut Pro (FCP) software (i.e. this has created progressive frames). Thus, the selected original reference for this present investigation consisted of 25 progressive frames per second (not 25 interlaced frames per second) and is compressed with MPEG-2.

Since compression performance is dependent on scene content, the various scenes selected from the bus dataset were characterised, classified and grouped based on local (just on the face) and global (on the entire sequence) scene properties. In total, 27 scenes were grouped, of which 2 were used for training the expert observers.

The training scenes were not included in the results. Scenes of 20 seconds duration were selected, to enable the temporal reduction processes of the video compression algorithms to adjust to the scene content. The local classification techniques discussed below focused on only 8 frames in scenes of 20 seconds duration. In this duration, a face that appeared in 8fps at an approximately consistent subject to camera distance, angle to the camera and under constant illumination was selected. The following paragraphs provide information on scene grouping (refer to Section 3.5.1 for information on the characterisation techniques).

1. Camera to subject distance. The average value among the 8 frames of the face was used to classify the face into a selected camera to subject distance group (see Table 4.1). The scenes were classified empirically into 2 groups:

close (44 pixels distance between the eyes, +/-4.5 pixels) and far (25 pixels distance between the eyes, +/- 3.5 pixels)

2. Scene lightness. The reference video footage was converted to a sequence of RGB TIFF still images and later to the CIELab space (see Section 3.3.4) before deriving the following L^* values. The average value among the 8 frames of the face was used to classify the face into a selected lightness group (see Table 4.1). The scenes fell into 5 groups of lightness using 2 types of illumination (daylight and bus illumination): 1) Medium (bus illumination): L^* 42 (+/- 11). 2) Medium (daylight): L^* 46 (+/- 6). 3) Low (daylight): L^* 8.5 (+/- 2.5). 4) High (daylight): L^* 92 (+/- 4.5). 5) Mixed (daylight): L^* 97 (+/- 2.5) and L^* 49.5 (+/- 15.5), (i.e. approximately half of the face had L^* 97.5 and the other half L^* 49.5).

   The medium skin lightness groups differ in terms of 'type' of illumination (i.e. bus illumination and daylight). It was observed that the camcorder produces noisier imagery under bus illumination at night than under daylight illumination. Daylight produces higher illumination levels than bus illumination. To compensate the exposure for decreased levels of illumination, when the bus lights are on, the camcorder increases the ISO settings resulting to increased noise levels. This is very likely to be the same for the actual CCTV systems installed in London buses. It was thus considered important to include bus illumination on its own in the investigation.

3. Angle of face to camera plane. 2 groups were derived: tilted angle and frontal angle (see Table 4.2).

4. Busyness. The grouping was made based on the measured spatial and temporal values only of the available 25 scenes. Their exact middle values (e.g. middle value from 2 to 5 is 3.5) were chosen as the limits. For example, the middle value for the spatial measures is 14.58 and for the temporal is 27.16 (see Table 4.2). The following 4 groups were created: 1) High Spatial ($>$ 14.58)-High Temporal ($>$ 27.16). 2) High Spatial ($>$ 14.58)-Low Temporal

(< 27.16). 3) Low Spatial (< 14.58)-High Temporal (> 27.16). And, 4) Low Spatial (< 14.58)-Low Temporal (> 27.16).

Figure 4.4 includes all 25 scenes used in the psychophysical investigations. Table 4.3 summarises the grouping of the scenes. For example, scene 1 (S1) belongs to the following groups: medium scene lightness (bus illumination), close camera to subject distance, frontal angle to the camera plane and low spatial-low temporal busyness.

Scene	**Camera to Subject Distance**			**Scene Lightness**		
Name	*Av.*	*std*	*Group*	*Av.*	*std*	*Group*
S1	41	2.03	Close	45	2.09	Medium (Bus)
S2	45	2.51	Close	45	2.20	Medium (Bus)
S3	40	2.43	Close	38	1.28	Medium (Bus)
S4	27	0.99	Far	31	0.69	Medium (Bus)
S5	28	1.28	Far	53	0.46	Medium (Bus)
S6	43	0.46	Close	52	0.68	Medium (Day)
S7	46	0.35	Close	54	0.96	Medium (Day)
S8	22	1.04	Far	47	2.20	Medium (Day)
S9	24	0.52	Far	48	0.83	Medium (Day)
S10	23	0.64	Far	40	0.95	Medium (Day)
S11	45	2.27	Close	6	0.73	Low (Day)
S12	48	1.96	Close	6	0.63	Low (Day)
S13	27	1.51	Far	11	2.15	Low (Day)
S14	29	0.64	Far	10	0.16	Low (Day)
S15	45	0.53	Close	88	0.57	High (Day)
S16	48	1.73	Close	94	0.63	High (Day)
S17	28	0.71	Far	90	0.78	High (Day)
S18	28	0.52	Far	97	0.40	High (Day)
S19	26	0.71	Far	96	0.49	High (Day)
S20	39	0.89	Close	65/95	6.12/2.16	Mixed (Day)
S21	45	2.51	Close	53/96	9.08/3.37	Mixed (Day)
S22	47	3.82	Close	34/98	4.67/1.22	Mixed (Day)
S23	25	1.30	Far	54/99	2.68/1.10	Mixed (Day)
S24	27	1.04	Far	46/100	2.03/0.04	Mixed (Day)
S25	26	0.76	Far	46/99	0.84/0.51	Mixed (Day)

Table 4.1: Scene measurements and grouping I. Derived measurements for the classification of each scene into groups. Where *Av.* is the mean measurement among the 8 frames, and *std* is the standard deviation and denotes the variance among the measurements in the 8 frames. The units for the average mean value for the camera to subject distance category is number of pixels. Whereas, for the scene lightness category is the L^* value. The group column indicates into which group each measurement has been classified.

Scene	**Angle to the camera**	**Scene Busyness**		
Name	*Group*	*Spatial*	*Temporal*	*Group*
S1	Frontal	11.12	22.92	Low Spa. Low Temp.
S2	Frontal	11.12	22.92	Low Spa. Low Temp.
S3	Frontal	11.12	22.92	Low Spa. Low Temp.
S4	Tilted	12.27	17.32	Low Spa. Low Temp.
S5	Tilted	10.95	29.01	Low Spa. High Temp.
S6	Frontal	13.97	29.08	Low Spa. High Temp.
S7	Frontal	14.48	15.66	Low Spa. Low Temp.
S8	Tilted	15.11	21.85	Low Spa. Low Temp.
S9	Frontal	17.47	25.74	High Spa. Low Temp.
S10	Tilted	16.51	31.45	High Spa. High Temp.
S11	Frontal	16.56	29.79	High Spa. High Temp.
S12	Frontal	18.21	32.41	High Spa. High Temp.
S13	Frontal	16.56	30.67	High Spa. High Temp.
S14	Tilted	14.92	27.95	High Spa. High Temp.
S15	Tilted	18.14	26.29	High Spa. Low Temp.
S16	Frontal	16.81	35.71	High Spa. High Temp.
S17	Frontal	17.20	36.20	High Spa. High Temp.
S18	Tilted	17.17	22.68	High Spa. Low Temp.
S19	Tilted	14.66	33.17	High Spa. High Temp.
S20	Tilted	13.30	34.67	Low Spa. High Temp.
S21	Tilted	13.30	35.84	Low Spa. High Temp.
S22	Tilted	14.54	38.67	High Spa. High Temp.
S23	Frontal	15.43	16.01	Low Spa. Low Temp.
S24	Frontal	16.25	18.47	High Spa. Low Temp.
S25	Tilted	16.04	24.17	High Spa. Low Temp.

Table 4.2: Scene measurements and grouping II. Derived measurements for the classification of each scene into groups. The values under the Spatial and Temporal categories represent the maximum value derived from their measurements (refer to Section 3.5.1). The group column indicates into which group each measurement has been classified.

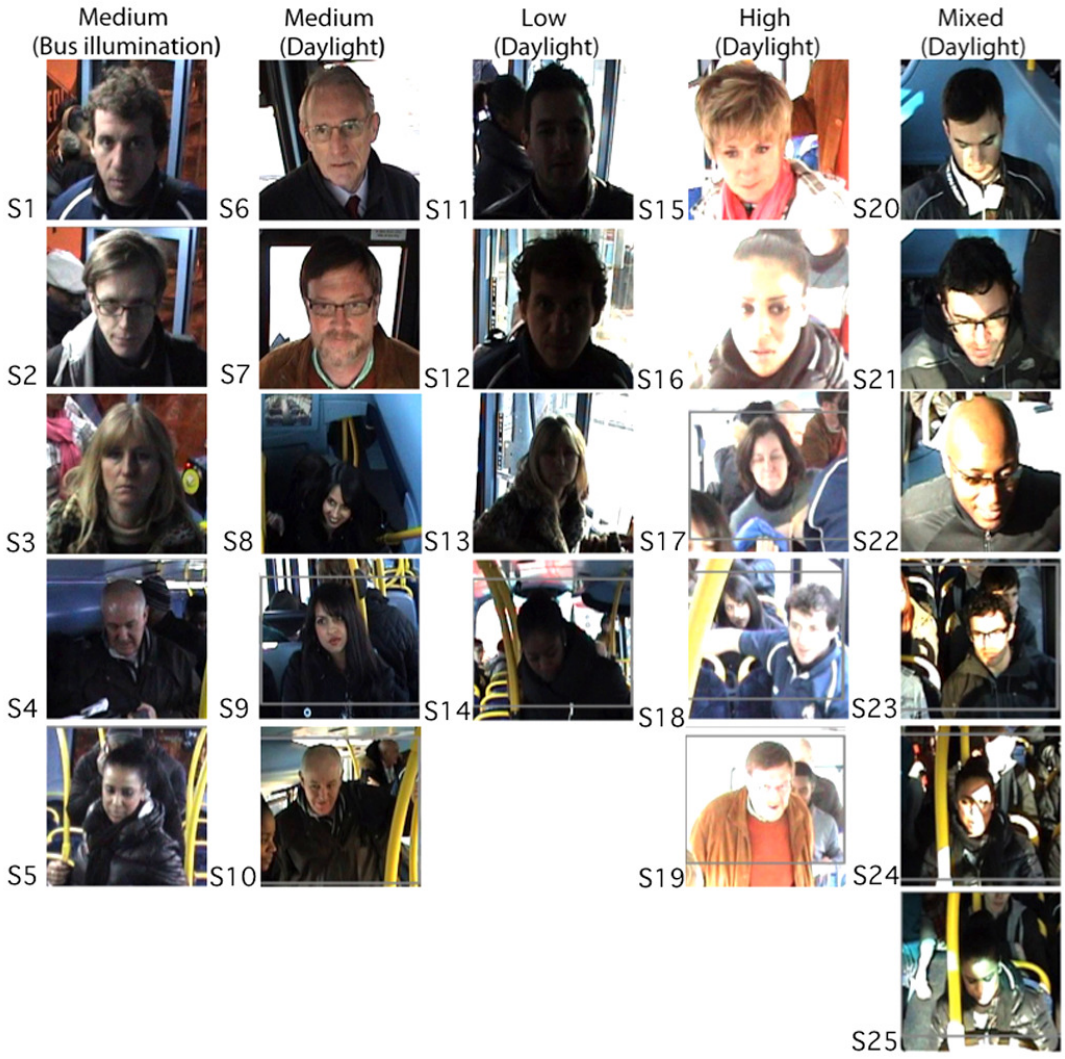


Figure 4.4: The 25 scenes under investigation. The 25 scenes grouped in columns based on the scene lightness property.

Group	Scene Name	Total
	Camera to Subject Distance	
Close	S1, S2, S3, S6, S7, S11, S12, S15, S16, S20, S21, S22	12
Far	S4, S5, S8, S9, S10, S13, S14, S17, S18, S19, S23, S24, S25	13
	Scene Lightness	
Medium (Bus)	S1, S2, S3, S4, S5	5
Medium (Day)	S6, S7, S8, S9, S10	5
Low (Day)	S11, S12, S13, S14	4
High (Day)	S15, S16, S17, S18, S19	5
Mixed (Day)	S20, S21, S22, S23, S24, S25	6
	Angle of Face to the Camera Plane	
Frontal	S1, S2, S3, S6, S7, S9, S11, S12, S13, S16, S17, S23, S24	13
Tilted	S4, S5, S8, S10, S14, S15, S18, S19, S20, S21, S22, S25	12
	Scene Busyness	
Low Spa. Low Temp.	S1, S2, S3, S4, S7	5
Low Spa. High Temp.	S5, S6, S20, S21	4
High Spa. Low Temp.	S8, S9, S15, S18, S23, S24, S25	7
High Spa. High Temp.	S10, S11, S12, S13, S14, S16, S17, S19, S22	9

Table 4.3: Summary of scene grouping. Each scene from figure 6 belongs to different groups. The totals indicate the total number of scenes in the specific group.

4.2.3 Identification of Key Scenes

The key scenes, those affected most by compression, were identified by carrying out a psychophysical investigation on the 25 grouped scenes. The MPEG Streamclip implementation encoder was employed to compress the scenes at selected target bitrates, using the video coding standard H.264/MPEG-4 AVC. Implementation encoders such as verification models used for compliance testing (e.g. Joint Model (JM) and FFpmeg) are often used by the scientific community; they allow the setting of over 50 parameters, such as quantisation parameters, I, P and B frames and the target bitrate. These verification models, when tuned properly, tend to apply 'high quality' compression, whilst encoders in the consumer and CCTV industry apply 'lower quality' compression [258]. It was decided that the verification models were not appropriate for this work. Thus, an encoder from the consumer industry was selected (MPEG Streamclip) with only bitrate control (i.e. no GOP size or B frames were selected), which complies with the security recording systems on buses.

Most of the scenes were compressed at 9 different bitrates, whilst some 'demanding' ones at 12 different bitrates, all at 25 frames per second. The 'demanding' ones were perceived to require less compression to maintain useful information than the rest of the scenes. The levels and ranges of compression were selected empirically, after careful visual examination, to provide enough data for the derivation of an accurate psychometric curve (see Section 3.4.3) [169]. The compression bitrates used were approximately the following in kilobits per second (kbps):

- 9 bitrates: 300, 400, 600, 800, 1000, 1200, 1400, 1600, 1800;
- 12 bitrates: 600, 800, 1000, 1200, 1400, 1600, 1800, 2000, 2200, 2400, 2600, 2800.

A similar method to the fidelity assessment (see Section 3.3.1) was implemented for the assessment of image usefulness of the compressed scenes. In the psychophysical experiment the observers were presented each time with 4 versions of the same scene, running simultaneously on a single carefully calibrated computer monitor. These were presented on a mid grey background, at 25 fields per second, as illustrated in Figure 4.5. The top left is the reference scene and the other 3 are compressed versions of the reference scene. Although the compression was applied on a 20 second scene at 25 frames per second, the observers were only presented with 8 frames, in which the face was placed within a grey square. The observers could see the displayed compressed version frame by frame and as many times as they wished before making their judgement.

During the experimental period, the EIZO CG210 LCD (pixel resolution: 1600H $\times$ 1200V) monitor was utilised and calibrated daily using the GretagMacbeth Eye-One Pro system. The monitor was calibrated to a white point D65 (6500K), at a luminance of $120cd/m^2$ (i.e. maximum brightness) using an sRGB ICC profile. Refer to Appendix A for further information in relation to the utilised monitor concerning it's calibration, transfer function, spatial uniformity, temporal instability

Figure 4.5: Example of the test display used in the identification of the key scenes. The top left scene is the reference and the remaining 3 scenes are compressed versions of the reference.

and viewing angle characteristics. Based on our current knowledge, there are no standards directly applicable to monitors used for CCTV viewing purposes. The experiment was conducted in dark conditions to minimise reflections and monitor flare. The specialist observers were asked to wear glasses, if they would normally do so in front of a monitor. Additionally, the observers were checked for colour deficiency with the Ishihara colour test for colour deficiency [259]; all of them had normal colour vision.

The observers consisted of 7 Metropolitan Police Service (MPS) police officers, 10 MPS surveillance officers, and 10 bus analysts. Table 4.4 provides a summary of the observers' average years of experience and purpose of use of security imagery.

	Bus Analysts	**MPS Police**	**MPS Surveillance**
Average years of experience	5 years	9 years	18 years
Use of security imagery	To recognise for security purposes and bus issues, gathering evidence for the police.	To recognise and provide evidence to court mainly for volume crime (e.g. antisocial behaviour, assaults).	To monitor activities and behaviours, recognise and provide evidence to court mainly for major crime (e.g. murder).

Table 4.4: Observers' background.

Instructions to the observers were given via a demonstration of a selected scene from the training set. The training scenes were excluded from the results. The instructions were:"The reference represents the maximum facial information that can be captured under the available illumination conditions and should be considered to have acceptable image usefulness. The aim is to find how much degradation (compression) from the reference is acceptable. You are required to respond with a *yes* or *no* to the question: Is the compressed version as useful as the reference in terms of facial information? You are judging only the face within the grey square, not the clothes or the surrounding area. Everything else that surrounds the face is irrelevant and should not influence your judgement. This experiment will help to identify the maximum acceptable degradation (compression) from an uncompressed reference. If you are paired while doing the experiment, you are allowed to discuss your thoughts with your partner, but your final answer should be independent of your partner's answer. Be aware of peer pressure. If you get bored or tired during the experiment, please inform the experimenter". In most cases, the observers were paired during the experiment. This is usual practice during police examination of CCTV footage.

The *yes/no* tasks have the property of being 'criteria dependent' [260]. For example, the observer might adopt his/her own criteria on the strength of the signal (facial information) before a *yes* response is obtained. If the criterion is loose, then a weak signal might be sufficient, whereas if a strict criterion is adapted then a relatively

strong signal might be required to obtain a positive response. The observers in this investigation have probably used criteria that have been derived from their individual work experience. This was not asked from the observers, they were only provided with the above mentioned instructions. Results are presented in Section 4.3 .

4.2.4 Testing of CCTV Systems

The identified key scenes from Section 4.2.3 were given to five suppliers of CCTV recording systems together with instructions on amount and ranges of compression. The key scenes were compressed at 4 frames per second (which was the requirement by TfL) and the compressed bitrates, in kbps were: 10, 160, 352, 544, 736, 928, 1120, 1312, 1504. The amount and ranges of compression were selected empirically, after careful visual examinations and observation of results obtained from the first psychophysical investigation. Each second consists of 25 frames. Reducing the frames from 25 to 4 per second has resulted, in the majority of cases, in an output from most CCTV recorders of 1 frame from the 8 frames with the face.

In this second psychophysical experiment, the methodology detailed in Section 4.2.3 was followed aside from 2 modifications: i) the mode of presentation of the experiment (see Figure 4.6), and ii) the number of observers involved. It was noticed in the previous psychophysical investigation that the observers would occasionally pay attention to the surrounding areas in the image of the faces. This possibility was eliminated, by cropping the surrounding areas. The number of observers was reduced to 2 MPS police officers and 9 bus analysts. All observers were trained on the task by participating in the first psychophysical investigation. The number of observers is still acceptable for fitting psychometric curves to their responses (see Section 4.3.1).

The observers had to judge the output of each CCTV recorder (1 frame) against the reference (8 frames) for each key scene. The performance of the five CCTV recording systems using the key scenes is presented in Section 4.4.

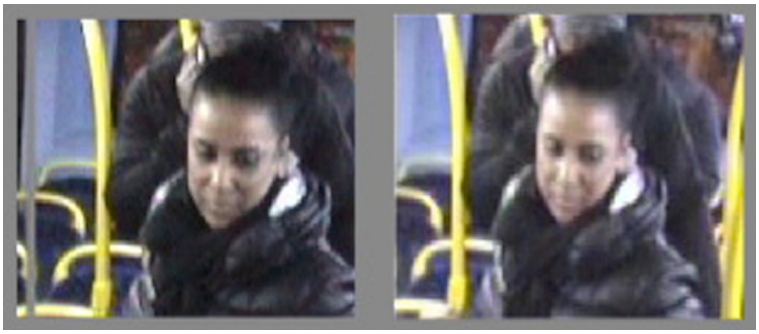


Figure 4.6: Example of the test display used in the testing of the CCTV systems. The left image is the reference whereas the right image is the compressed version of the reference.

4.3 Results from the Identification of Key Scenes

The data obtained from the first psychophysical investigation were modelled by fitting psychometric curves (see Section 4.3.1), as described in [261], for each scene. The curve describes the response of the observers' sensory mechanism to the different stimulus levels (i.e. compression levels). The sigmoid logistic function was fitted to the obtained psychophysical data points (i.e. proportion of *yes* responses) at each different level of compression in kbps. The logistic function is given as:

$$F(x) = (1 - \Lambda)\frac{1}{1 + \exp(-\beta(x - \alpha)} \tag{4.1}$$

The shape of the curve is established from parameters x, α, β, and Λ. x corresponds to the compression bitrate (e.g.100kbps); α corresponds to the absolute threshold (i.e it is at the point of 50% *yes* responses); β to the gradient of the curve; Λ is the stimulus independent lapse rate and was fixed for most fittings at 0.01 except for scene number 24 were the value was fixed at 0.02 (i.e. it produced a more acceptable fit to the data points). The lapse rate parameter determines the upper bound of the curve given by $(1 - \Lambda)$ (see Eq. 4.1). Observer lapses need to be taken into consideration as they can introduce biases to the estimated α and β parameters. The effect of lapses can be minimised by setting it to a small but non-zero value, such as 0.01 or 0.02 [262, 263]. The maximum-likelihood estimation technique was used to estimate the curve parameters α and β [264, 265].

Figures 4.7 and 4.8 present the obtained curves from the first psychophysical experiment. Tables 4.5 and 4.6 include measures of the obtained curves: a) the estimated curve parameters α and β, b) the α and β estimated standard errors (SE), c) the value of goodness of fit (pDev), and d) the value, in kbps, that corresponds to the 75% proportion of observers *yes* responses.

The α and β parameters are just estimates of the true parameters of the sensory mechanism. The errors on the estimated parameters were derived by implementing a non-parametric bootstrap analysis, which is a Monte Carlo re-sampling technique. Bootstrap methods produce simulated repetitions using the data from the original experiment [266, 267]. The standard deviation among the obtained values from the simulated experiments is used as the measure for standard error. In this investigation the recommended 400 converged simulated experiments were used in order to obtain the errors [263]. Stimulations that did not converge were excluded. A parametric method is most frequently suggested when the psychometric curve is a good fit to the data points [263]. A non-parametric method was employed in order to sustain a harmonised analysis among all the fitted curves, the good ones and the less good ones. Additionally, there is a controversy on which of the 2 methods produces better error estimates [224].

The goodness of fit is a measure that describes how well the curve fits the data. The measure derives the pDev value (i.e. is the statistical p-value) that ranges between 0 (a bad fit) to 1 (the best fit). When the pDev value is less than 0.05 then the fit is considered unacceptably poor. When the curve falls precisely on the points then this indicates a good fit. The goodness of fit measure was calculated using 400 bootstrap simulated experiments and the method is illustrated in [262]. Both α and β were set as free parameters and Λ as a fixed parameter during the process of estimating the errors and the pDev values. The 75% of *yes* responses was taken as the just noticeable difference (JND) point on the psychometric curve to identify the acceptable bitrates for the scenes under test. It is typically the value used in qualitative work relating to imaging science [169, 268]. The 50% is defined

as the absolute threshold. This is the point where the observers are starting to seeing, in this case, the compressed version to be equal in terms of usefulness to the reference [243]. The subsequent sections provide an analysis of the results from the first psychophysical investigation.

4.3.1 Psychometric curve fitting

Figure 4.7 and Table 4.5 present the results derived from fitting curves to the data, for each observer group. Figure 4.8 and Table 4.6 present the results from the 25 scenes under investigation. The errors on the obtained β parameters were greater than for the α parameters. Error values could be reduced by increasing the number of observers and, also, by having a better distribution of stimulus intensities. In image related investigations, it is recommended to use between 10 to 30 observers. The use of more observers will increase the precision of the estimated values (decrease the error) and not their accuracy [238]. The error estimates in this work are included only for references. The fitting results have shown the pDev values to score above 0.05 for all the different types of observers (see Tables 4.5) and for each of the 25 scenes under test (see Table 4.6) and thus all the fitted models/curves are acceptable. Also, common goodness of fit methods used by linear models such as R-Squared have been found inadequate for non-linear models [269]. A psychometric function is a non-linear model. For this reason error bars have not been applied to the obtained fitted models, instead the models with their associated errors on the parameters (and pDev value) have been utilised for the analysis/explanation of results.The parameters of a psychometric function (or model) are what it has been estimated from the raw data.

Group	α	SE	β	SE	pDev	75% (kbps)
$Bus_{Analysts}$	2.804	0.042	7.736	2.429	0.977	894
MPS_{Police}	2.840	0.048	6.888	2.516	0.950	1014
$MPS_{Surveillance}$	2.661	0.038	11.308	3.400	0.923	578

Table 4.5: Data from curve fitting for each observer group illustrating the parameter estimates along with their estimated standard error (SE) for the different groups of observers. The goodness of fit is given by the pDev value. The 75% of *yes* responses, in kbps, for each curve is also provided.

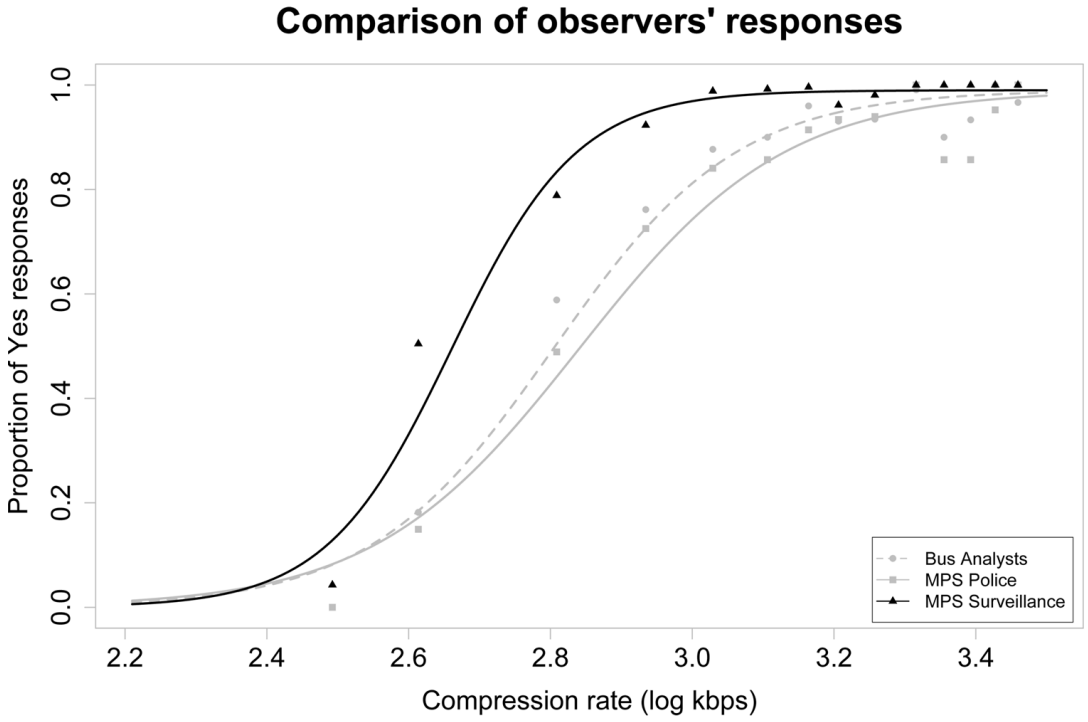


Figure 4.7: Psychometric curves for each observer group. The mean value of all the tested scenes was used for each observer group.

When the calculated coefficients (+/- standard error-SE) between models overlap (for α and β coefficients) then this is an indication of there being no difference between the investigated models. For example, the estimated model coefficients (see Table 4.6) between bus analysts (α=2.804 +/- 0.042, β=7.736 +/- 2.429) and police officers (α=2.840 +/- 0.048, β=6.888 +/- 2.516) overlap indicating that there is not difference between these 2 models. These 2 models differ with the model obtained from the surveillance officers (α=2.661 +/- 0.038, β=11.308 +/- 3400). The police officers have tolerated less compression for maintaining usefulness, from the original reference, than the bus analysts and surveillance officers (see Figure 4.7). The point at 75% of *yes* responses for the police officers is at 1014kbps, for

the bus analysts at 894kbps, and for the surveillance officers at 578kbps (see Table 4.5). The bus analysts are considered as having the highest technical understanding of video compression and video systems, followed up by the surveillance officers and last the police officers. The surveillance officers have started as police officers and their work, in most cases, involves monitoring (such as following and recording) targeted individuals and gathering evidence to present in court or to help with a case. Their experience and in general the use of different sources of information of the targeted individual (e.g. knowing where the individual has been helps to identify the correct CCTV system to extract supportive imagery) make even a highly compressed CCTV imagery usable for the completion of their task. This is not the case for the police officers as the individuals are often unknown and thus making a recognition task from facial imagery more difficult. A couple of surveillance officers where around the age of 50 years old (noise in the human visual system might have affected the results) but most of the observers were younger (between 25 to 40 years old).

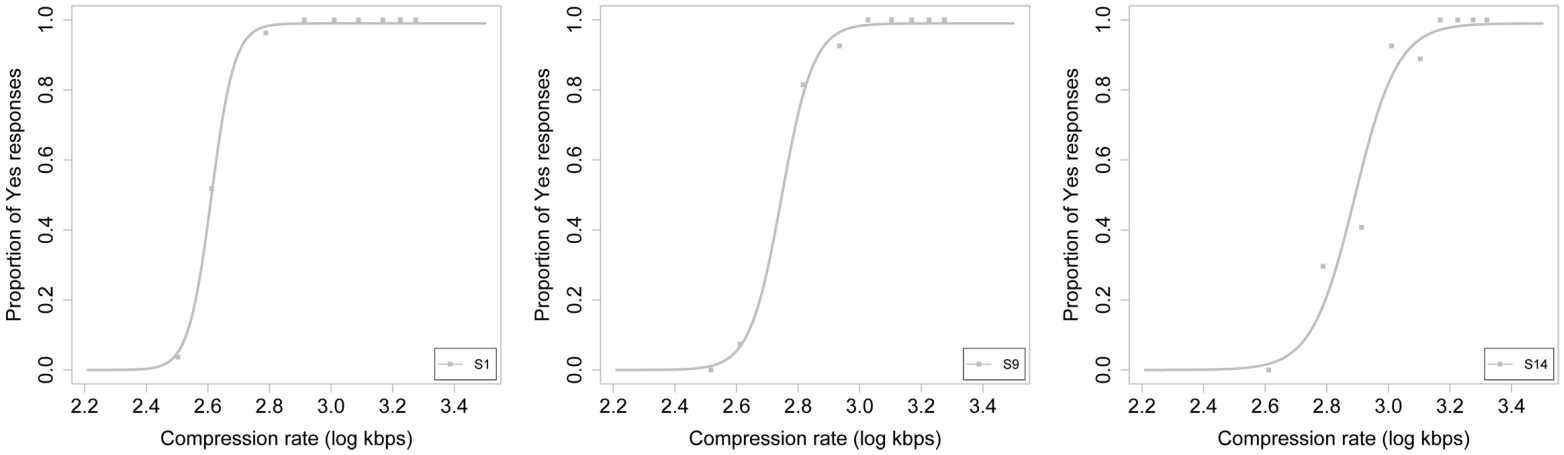


Figure 4.8: 3 example psychometric fitted curves, for scenes S1, S9 and S14. With high, medium and low derived pDev values respectively.

Scene	α	SE	β	SE	pDev	75% (kbps)
S1	2.611	0.016	26.917	5.699	0.930	450
S2	2.640	0.020	13.553	2.234	0.135	529
S3	2.721	0.022	16.374	2.904	0.730	617
S4	2.683	0.020	17.017	2.663	0.223	562
S5	2.795	0.019	13.901	2.959	0.868	753
S6	2.659	0.021	16.199	3.127	0.150	536
S7	2.530	0.000	524.084	0.000	0.845	340
S8	2.531	0.000	531.635	0.000	0.198	341
S9	2.747	0.021	19.426	5.724	0.673	639
S10	2.856	0.015	21.501	5.318	0.798	811
S11	2.953	0.018	15.565	2.421	0.715	1063
S12	3.057	0.018	10.441	1.300	0.863	1467
S13	3.044	0.019	9.252	1.063	0.090	1469
S14	2.891	0.018	14.359	1.917	0.058	934
S15	2.677	0.018	22.214	5.147	0.925	534
S16	2.659	0.017	25.929	5.302	0.388	505
S17	3.079	0.014	14.522	2.241	0.305	1437
S18	2.709	0.021	15.07	1.897	0.145	609
S19	2.828	0.016	17.881	3.744	0.153	780
S20	2.634	0.022	15.346	12.334	0.103	510
S21	2.728	0.021	17.246	4.836	0.978	623
S22	2.709	0.028	13.1	1.925	0.423	625
S23	2.662	0.020	17.228	3.590	0.068	535
S24	2.665	0.018	19.923	12.208	0.063	530
S25	2.825	0.017	15.877	2.476	0.553	788

Table 4.6: Curve fitting data for each of the 25 scenes. The parameter estimates along with their estimated standard error (SE) for each of the 25 scenes under investigation. The goodness of fit is given by the pDev value. The 75% of *yes* responses, in kbps, for each curve is also provided.

Few of the scenes have fallen under the exact combination in terms of camera to subject distance, scene lightness and so on (see Table 4.3). For example, scenes S1 (75% =450), S2 (75%=529), and S3 (75%=617) represent one exact combination, where 75% is the proportion of *yes* responses (see Table 4.6). Another exact combination is for scenes S11 (75%=1063) and S12 (75%=1467). One more is for scenes S20 (75%=510) and S21 (75%=623). Lastly, another one is for scenes S23 (75%=535) and S24 (75%=530). Most of the exact combinations have produced similar results except for S11 and S12, where the difference is more than 300kbps. Furthermore, in the high lightness group of scenes (S15, S16, S17, S18, and S19) only S17 was affected to a greater degree by compression than the rest. The presented scene characterisation and classification methods might not be enough in order to describe

the properties of the scenes/faces. For example, the results could be influenced by the degree of distinctiveness/uniqueness or overall appearance of the actual faces in the scenes. Distinctive faces are more memorable [270]. Bruce et al have found that distinctiveness correlates with how much a face deviates from an 'average face' [271]. Perhaps, distinctive faces (e.g. Arnold Schwarzenegger) can take more compression than typical faces (e.g. Leonardo DiCaprio) [270]. Furthermore, Penry provides guidance on how facial features/shapes can be classified [272].

4.3.2 Comparison of the classified scene groups

The point at 75% (in kbps) of *yes* responses for each of the 25 scenes was chosen for further analysis. This analysis investigates the differences and similarities between the classified scene groups. Table 4.7 illustrates the descriptive statistics for each scene group. Mostly, the statistics describe the variability of the obtained values of the scenes in each scene group. The values of the mean and the median for each scene group are similar, a result indicating near normal distributions. Although, parametric statistics are used with normal distribution, in the following analysis a non-parametric method was used due to the small number of scenes in each group.

Table 4.8 shows the results from the Wilcoxon Rank Sum Test [273]. This is a non-parametric test that ranks the values of 2 independent samples and compares the differences between the 2 rank totals. This method focuses on the median rather than the mean. It derives the p statistical value at 0.05 significance level, below which 2 groups will be considered as statistically different. This method allows gathering the similar groups into a single one.

Table 4.8 reveals the similarity/difference between each scene grouped category. For example, when 2 groups are similar then they could be further classified to the same group (e.g. no significant difference between 'medium lightness-bus illumination' and 'medium lightness-daylight' scene groups).

Group Name	N	Range	Min	Max	Mean	Median	std
Scene Lightness							
MED (BUS)	5	303	451	753	582	562	113
MED (DAY)	5	471	340	811	533	536	202
LOW (DAY)	4	534	934	1469	1233	1265	276
HIGH (DAY)	5	933	505	1437	773	609	386
MIXED (DAY)	6	278	510	788	602	579	104
Camera to Subject Distance							
CLOSE	12	1127	340	1467	652	534	325
FAR	13	1127	341 1	469	784	753	335
Angle of Face to the Camera							
FRONTAL	13	1128	340	1469	778	536	421
TILTED	12	593	341	934	656	624	163
Scene Busyness							
HIGH SPA. HIGH TEMP.	9	964	505	1469	1010	934	372
LOW SPA. LOW TEMP.	5	277	340	617	500	529	108
HIGH SPA. LOW TEMP.	7	447	341	788	568	535	135
LOW SPA. HIGH TEMP.	4	243	510	753	606	579	110

Table 4.7: Descriptive statistics at 75% of *yes* responses. Where N is the number of scenes in the group. Range is the difference between the minimum (MIN) and maximum (MAX) values. The range, mean, median and standard deviation (STD) are measures of variability of the obtained values of the scenes in the group.

The results have shown that the 'low-daylight' lightness group is significantly different from all the other lightness groups except for group 'high-daylight', which it is marginally significant. The groups 'medium-bus illumination', 'medium-daylight', 'high-daylight' and 'mixed-daylight' can be further classified to the same group as there is not a significance difference among them. The 'low-daylight' scenes were affected more by compression than the rest of the lightness groups as the mean value of the scenes for the 75% of *yes* responses is at 1233kbps where for the rest of the lightness scenes is less than 773kbps (see Table 4.7).

There is marginally significant difference between the 2 camera to subject distance groups (Table 4.8) were scenes in the far distance group (mean value of 75% *yes* responses at 784kbps) were affected more by compression than the close distance group (mean value at 75% *yes* responses at 652kbps-see Table 4.7).

(I)Group	(J)Group	Mean Difference	p	h
Scene Lightness				
MED (BUS)	MED (DAY)	49	0.841	0
	LOW (DAY)	651	0.016	1*
	HIGH (DAY)	191	0.548	0
	MIXED (DAY)	19	0.662	0
MED (DAY)	LOW (DAY)	700	0.016	1*
	HIGH (DAY)	240	0.548	0
	MIXED (DAY)	68	0.931	0
LOW (DAY)	HIGH (DAY)	460	0.063	0*
	MIXED (DAY)	631	0.009	1*
HIGH (DAY)	MIXED (DAY)	171	0.931	0
Camera to Subject Distance				
CLOSE	FAR	134	0.097	0*
Angle of Face to the Camera				
FRONTAL	TILTED	122	0.765	0
Scene Busyness				
HIGH SPA. HIGH TEMP.	LOW SPA. LOW TEMP.	510	0.007	1 *
	HIGH SPA. LOW TEMP.	442	0.016	1*
	LOW SPA. HIGH TEMP.	405	0.050	0*
LOW SPA. LOW TEMP.	HIGH SPA. LOW TEMP.	68	0.343	0
	LOW SPA. HIGH TEMP.	106	0.286	0
HIGH SPA. LOW TEMP.	LOW SPA. HIGH TEMP.	38	0.788	0

Table 4.8: Wilcoxon Rank Sum Test. The (I) group is compared against the (J) group. When the p value is less then 0.05 then the groups are significantly different. Significantly different groups have scored 1 in the h column and marked with an asterisk. The 0 values in the h column that have been marked with an asterisk are results that are marginally significant.

There was not a significance difference between the 2 angle of face to camera plane groups so they could be further classified to the same group. This requires a further investigation with perhaps higher degrees of tilted angles.

The busyness of the scenes affected compression performance. Scenes with 'high spatial-high temporal' busyness were significantly different from all the other busyness groups, except for group 'low spatial-high temporal' which it is marginally significant (see Table 4.8). All the busyness groups excluding the group of 'high spatial-high temporal' can be classified into one group. The scenes in the 'high spatial-high temporal' group have given a mean value of 1010kbps for the 75% *yes* responses whereas for the other groups it is around 550kbps (see Table 4.7). This

result is expected as a high busyness scene, due the the high information content both spatially and temporarily, will require higher bitrates in comparison to a low busyness scene (i.e. it entails less spatio-temporal information) to maintain the same information from an 'uncompressed' reference.

4.4 Results from Testing of the CCTV Systems with the Selected Key Scenes

One scene from each of the following 4 scene lightness groups was selected: 'high-daylight', 'medium-daylight', 'medium-bus illumination' and 'mixed-daylight'. A further 2 scenes from the 'low-daylight' group were selected. All 6 scenes, were these most affected by the compression. These key scenes (illustrated in Figure 4.9) were given to the CCTV suppliers for further investigation of the acceptable compression bitrates on London buses.

Figure 4.9: The selected key scenes.

Figure 4.10 illustrates an example of the output of the CCTV systems (labelled A, B, C, D and E) for key scene S12. As mentioned above, in most cases the CCTV systems exported, 1 frame from the 8 reference frames of the face. Even a small changeability in terms of subject to camera distance within each individual scene has affected the obtained results. For example, system C in Figure 4.10 has obtained more *yes* responses at 736kbps than at 1120kbps because at 736kbps the face is closer to the camera. This could have been completely controlled by using still images, but it would not have replicated reality.

Additionally, Figure 4.10 illustrates an example of the visual differences between the outputted images from each CCTV system. System C has brightened the scene (by enhancement) whereas compression artefacts are more visible for systems D and

E. The systems are behaving differently even though all of them are based on the H.264/MPEG-4 AVC compression standard. This presents challenges in drawing conclusions about universal 'average' bitrates.

The results from the second psychophysical investigation illustrate the unpredictable nature of CCTV recording systems. For example, by reducing the frame rate from 25 to 4 has outputted 1 image from the 8 images of the face. This outputted 1 image might be the worst, or the best-case scenario from the 8 available images of the face. Even a slight difference in terms of camera to subject distance within each individual scene has been shown to affect the results for the CCTV systems. For this reason, the analysis of the results is based on the performance of all five CCTV recording systems for each key scene. The same curve fitting method from the first psychophysical experiment was applied. 3 curves were fitted for each key scene: a) the worst performance to the minimum points (lowest fit), b) the middle performance to the average points (average fit), and c) the best performance to the maximum points (highest fit). The lapse rate (Λ) was fixed for most fittings at 0.01 except for S12 highest fit were the value was fixed at 0.02 (i.e. produced a more acceptable fit to the data points).

Figures 4.11 and Table 4.9 present the results obtained from the testing of the CCTV systems. The application (linked to TfL) was seeking the absolute minimum bitrate to maximise data storage, so a 60% of observers yes responses was recommended to be used on London buses, which is higher then the absolute threshold of 50%.

Figure 4.10: An example of the outputs of the CCTV systems (labelled A, B, C, D and E) for key scene 12. The images on the top row are the 8 images of the reference. The second row shows the exported images from system A at different kbps (e.g. between 10 1504 kbps). The third row shows the exported images from System B and so on.

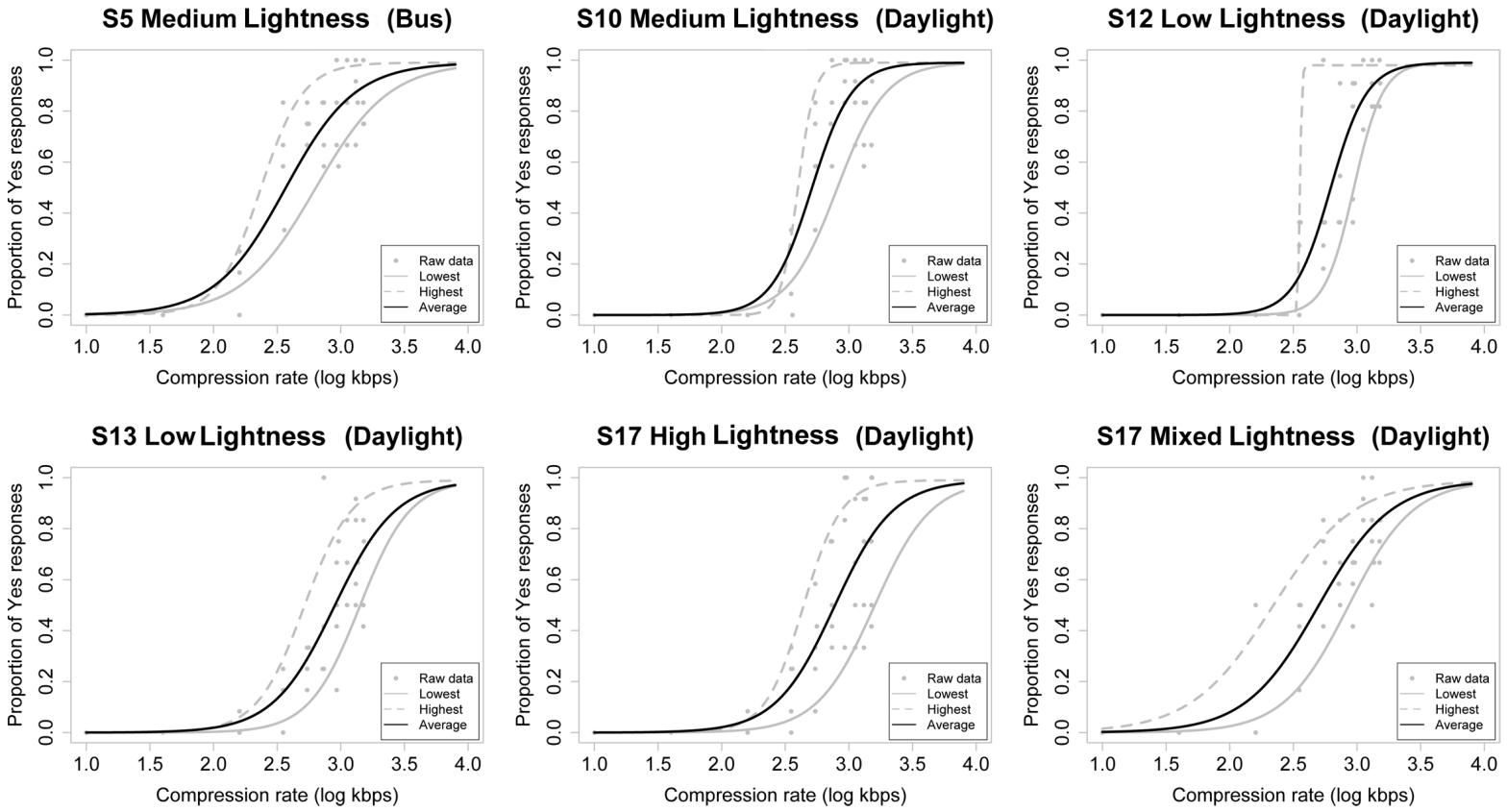


Figure 4.11: Psychometric curves for each key scene. 3 curves were fitted to all the data points derived from all the systems for each scene: a) worst performance curve using the lowest points (lowest fit) b) average performance curve using the average points (average fit), and c) best performance curve using the highest points (highest fit).

Scene Name	α	SE	β	SE	pDev	60% (kbps)
$S5_{LOWEST}$	2.799	0.063	3.434	0.709	0.495	840
$S5_{AVERAGE}$	2.563	0.075	3.639	0.822	0.898	480
$S5_{HIGHEST}$	2.37	0.076	5.958	4.513	0.510	277
$S10_{LOWEST}$	2.906	0.047	5.216	1.033	0.083	975
$S10_{AVERAGE}$	2.707	0.042	6.327	1.829	0.653	596
$S10_{HIGHEST}$	2.599	0.03 1	5.27	8.094	0.640	423
$S12_{LOWEST}$	2.974	0.031	8.717	1.982	0.805	1055
$S12_{AVERAGE}$	2.792	0.042	6.959	1.504	0.895	714
$S12_{HIGHEST}$	2.552	0.001	127.817	6.156	0.174	357
$S13_{LOWEST}$	3.149	0.066	1.588	1.588	0.650	1716
$S13_{AVERAGE}$	2.949	0.062	4.14 0	1.038	1.00	1131
$S13_{HIGHEST}$	2.713	0.052	5.356	1.443	0.090	621
$S17_{LOWEST}$	3.199	0.092	4.444	1.175	0.725	1977
$S17_{AVERAGE}$	2.887	0.055	4.324	1.040	0.950	969
$S17_{HIGHEST}$	2.642	0.056	6.65 0	1.77	0.685	509
$S25_{LOWEST}$	2.932	0.059	3.923	0.862	0.348	1100
$S25_{AVERAGE}$	2.695	0.068	3.502	0.827	0.808	657
$S25_{HIGHEST}$	2.338	0.107	3.114	0.656	0.273	299

Table 4.9: Data from curve fitting for each key scene from the CCTV systems illustrating the parameter estimates along with their estimated standard error (SE) for each curve. The goodness of fit is given by the pDev value. The 60% of *yes* responses, in kbps, for each curve is also provided.

Based on the data analysis in Table 4.9, results and recommendations were provided to TfL. It was recommended that, during daytime, when there is variable illumination, to set the bitrate to approximately 1977kbps (derived from the worst performance curves, scene S17) and during night-time, when the bus illumination is on, to reset the bitrate to around 840kbps (derived from the worst performance curve for constant bus illumination, scene S5).

4.5 Comparison between CCTV and Industry

Table 4.10, shows a comparison between the results from the consumer industry compressor (MPEG Streamclip-SC) in the first investigation and from the CCTV systems in the second investigation at 60% of *yes* responses for each key scene. This comparison helps to understand the performance of CCTV recording systems and thus employ appropriate testing methods for such systems.

	Scene Name					
	S5	*S10*	*S12*	*S13*	*S17*	*S25*
Industry (SC)	670	752	1255	1231	1285	711
$CCTV_{LowestFit}$	840	975	1055	1716	1977	1100
$CCTV_{AverageFit}$	480	596	714	1131	969	657
$CCTV_{HighestFit}$	277	423	357	621	509	299

Table 4.10: A comparison between CCTV and Industry compressors at 60% of *yes* responses for each key scene.

The performance of the consumer industry compression at 60% of *yes* responses is in most cases in the middle between the worst (lowest fit) and average (average fit) values of the CCTV systems. Also, the CCTV systems for all the fits have performed better than the consumer industry compression for scene 12 (required less bitrate to maintain facial information). This is because the CCTV systems have enhanced the dark areas by making them look brighter and thus revealed more information within the image. Additionally, it is observed that the CCTV systems might have performed some sharpening to the images, as a result making the information more visible. This does not mean that the image itself will have more information than the consumer industry compressed version. For example, the highest and average curve fits of the CCTV systems outperformed the consumer industry compressor by requiring less bitrate.

4.6 Discussion

Acceptable bitrates for video compression depend largely upon scene content properties. Dark scenes, far distance scenes and scenes with high levels of spatial-temporal busyness were found the most challenging to compress, requiring higher bitrates to maintain useful facial information, necessary for face recognition. Each of the aforementioned groups was made most susceptible to compression either from the scene property itself (i.e. dark and far distance scenes) or from the way the compression algorithm works with the scene content (i.e. high levels of spatio-temporal busyness scenes). Additionally, the dark and far distance scenes could be

considered as having less useful facial information in the reference (in comparison to the medium lightness and close distance groups) and they were affected more by compression than the rest of the groups.

The findings of this study can be easily extended to others applications. Scene content classification and grouping allows the use of the video dataset (and the selected scenes under investigation) to be valid for other CCTV applications at large. Furthermore, the classification provides exact knowledge on the scene properties that compression algorithms have been assessed with and thus the circumstances of data collection become irrelevant. For example, low lightness scenes are affected more by compression than medium lightness scenes, and this is valid not solely for CCTV but for any human face recognition application. This is because a low lightness scene has a lower signal-to-noise ratio and entails less visual information than a medium lightness scene and therefore is more susceptible to compression for human visual tasks.

CHAPTER 5

Case study 2: Comparative performance between human and automated face recognition systems, using CCTV imagery, different compression levels and scene properties

Automated face recognition systems should be aiming to work efficiently with CCTV imagery obtained under uncontrollable environmental conditions and compression. This chapter is a continuation of the previous human investigation in Chapter 4. Results and test material obtained from the human investigation are also utilised here. The aim of this chapter is to identify relationships between human and automated face recognition systems with respect to compression. Further, to identify and compare the most influential scene properties on the performance of each recognition system. Findings have the potential to broaden the methods used for testing automated face recognition systems and thus improve their performance.

5.1 Introduction

As it has been mentioned in Section 2.1.2, automated face recognition systems are normally assessed using large datasets whereas individual and unique scene content characteristics are not taken into consideration. Aggarwal *et al.* [42] have realised that performance of face recognition systems differs among the available datasets because of the dissimilar captured facial properties under each dataset. At the moment there is little research on the effects of individual scene content characteristics on the performance of automated systems (e.g. face recognition and analytics systems). Nevertheless, the term image quality is utilised in the same manner between automated systems and human operatives (see Section 3.3.2) [16, 18, 274]. Image quality for security systems (i.e. both automated and human) relates to the suitability of the imagery to satisfy, in this case, a face recognition task.

Results from the human investigation in Chapter 4 have shown that under-exposed scenes (dark scenes), far camera to subject distance scenes and scenes with high spatio-temporal busyness information were the most challenging to compress, requiring higher bitrates to maintain useful facial information. Overall, the correctly-exposed scenes entailed visually more facial information and were perceived to be immune to compression than under-exposed scenes. Thus, compression in human face recognition has less effect on correctly-exposed scenes.

A number of studies have shown that compression does not adversely affect the performance of automated face recognition systems (see Section 2.1.2) [34–37]. These findings were derived by analysing the results based on correct recognition rate. In this present work the results are analysed using a distance measure between a degraded image from its reference version, which complies with the methodology utilised for the human investigation in Chapter 4. Further, this will allow a direct comparison of the results obtained between the human and automated face recognition investigations.

Results and test material obtained from the human investigation in Chapter 4 are also used here. In the present investigation 4 systems are assessed, 1 human face recognition (HFR) system, and 3 basic automated face recognition (AFR) systems [51]: a) Principal Component Analysis (PCA), b) Linear Discriminant Analysis (LDA), and c) Kernel Fisher Analysis (KFA). Refer to Section 2.1.2 for a detailed description of automated face recognition algorithms.

In summary, this chapter provides information on testing the aforementioned face recognition systems with 'uncompressed' (i.e. the reference) and compressed footage (i.e. 25 scenes compressed with H.264/MPEG-4 AVC video coding standard using 2 types of encoders) consisting of quantified scene (footage) properties. These include measures of camera to subject distance, angle of the face to camera plane, scene lightness, and spatio-temporal busyness. Results from the human investigation are analysed using a different approach from the one employed in Chapter 4 in order to sustain a systematic statistical analysis among all the face recognition systems (i.e. AFR and HFR systems). For example, modelling of data is implemented based on group properties (i.e. fitting of a single model to the scenes belonging to the low lightness group) rather than on individual scenes as in Chapter 4.

Section 5.2 presents the experimental methodology. Data analysis of the results is described in Section 5.3. Lastly, in Section 5.4, conclusions are drawn.

5.2 Methodology

The test material in this current investigation consist of different formats (reference and degraded) and implementations of H.264/MPEG-4 AVC. A more detailed representation of the test materially can be found in Section 4.2.2. The following points provide a summary of the material under investigation:

- *25 reference 'uncompressed' scenes.* The reference 'uncompressed' format was compressed using MPEG-2 at approximately 25Mbits/s, 25 frames per second (25fps), and 4:2:0 chroma subsampling. This compression was applied to

the original recorded footage (i.e. DV format) in order to provide the test scenes on a DVD to the CCTV suppliers for the testing of the CCTV systems on London buses. Empirical observations by the experimenter showed no visible difference between the original recorded (DV format) and the assigned reference (MPEG-2 format) footage.

- *25 scenes, compressed with MPEG Streamclip implementation.* The compression bitrates used were at 25fps, with video coding standard H.264/MPEG-4 AVC, and approximately the following in kilobits per second (kbps): 300, 400, 800, 1000, 1200, 1400, 1600, 1800 and 2000.

- *The 6 key scenes, compressed with 5 CCTV recording systems.* These were the most affected scenes as shown in section 4.4. The compression bitrates used were at 4fps, with video coding standard H.264/MPEG-4 AVC, and approximately the following in kbps: 10, 160, 352, 544, 736, 928, 1120, 1312 and 1504.

Overall 4 face recognition systems have been assessed with these test material, 1 human face recognition (HFR) system, and the 3 basic automated face recognition (AFR) systems mentioned in Section 5.1.

Similarly to the human investigation, the automated systems were assessed based on similarity score distance (between a degraded image from its reference version) and not on correct recognition rate. Similarity scores provide a distance measure of facial information between 2 images of faces, or biometric signatures [275, 276]. The following equation 5.1 [277] provides the Euclidean distance similarity measure that has been employed in this investigation.

$$||x - y||_e = \sqrt{|x_i - y_i|^2} \tag{5.1}$$

The equation 5.1 finds the minimum distance $||x-y||_e$ between the weighted vectors of the probe/unknown (x_i) and training (y_i) images. The 3 AFR systems under investigation were executed using a publicly available MATLAB face recognition

toolbox [51, 278, 279].

The testing of the AFR systems included the following actions:

1. *Normalisation of facial images.* The footage of all 25 scenes (reference and degraded) was converted, with the MPEG Streamclip software, into a sequence of colour still images in TIFF uncompressed format. In the case of the footage generated from the CCTV DVR systems, the manufacturers' software was used to export stills in most cases in TIFF format (some CCTV DVRs did not support TIFF but rather PNG or JPEG formats). As mentioned in Chapter 4, each scene included a consistent face that appeared in 8 images. These 8 images were used to extract only the facial regions, based on eye coordinates (i.e. these were the same among reference and degraded images with MPEG Streamclip encoder but not with the CCTV DVR encoders). CCTV DVRs have altered the size (i.e. by recording at a lower resolution) of the original reference scenes. This indicates that the reference and degraded images with MPEG Streamclip encoder were exactly the same in terms of selected facial region but not exactly the same as the ones degraded with CCTV DVR encoders.

   All the extracted facial regions were normalised in terms of geometry (i.e. orientation), size (i.e. rescaled to 125×125 pixels) and were saved as colour images, in TIFF format (see Figure 5.1). This has resulted in the normalisation of 8 images for each scene (single individual) and each type of footage (reference and degraded with MPEG Streamclip encoder). In case of the CCTV DVRs, 1 image for each scene was normalised. This is because the DVRs have recorded the reference footage at 4fps, instead of 25fps, and have outputted 1 face image from the 8 face images available in the reference.

Figure 5.1: Extraction (left) and normalisation (right) of facial region.

The rescaling process (i.e. to 125×125 pixels) of the extracted facial regions has smoothed out the original facial images; both compressed and reference. Figure 5.2 illustrates examples of 2 facial scenes, S6 and S9, entailing close and far distance to the camera properties respectively. Also, the appearance of them before and after the rescaling or normalisation process is illustrated. Additionally, Figure 5.2 provides visually the effect of rescaling for the reference, 800kbps, 400kbps and 300kbps footage types. Even the reference images (of scenes S6 and S9) appear to have been smoothed out on the edges after the rescaling process. The rescaling process is unavoidable as any face recognition system will need to normalise facial images (training and unknown) before comparing them with each other; the normalisation process eliminates extra unwanted variations (e.g. positioning of head and size) within a dataset of faces. It is obvious that the rescaling process has smoothed out facial information including both edges and compression artefacts. One way to avoid this smoothing out effect on the compression artefacts would have been to first apply the rescaling process and later the compression. This option was omitted as it does not replicate usual practice.

Even though the rescaling process has smoothed out facial information (i.e. edges and compression artefacts) it seems that it has not reduced facial information. For example, the same information is shown between the un-

normalised and their normalised versions for each type of footage (reference and compressed versions) for both scenes S6 and S9. Perhaps the presence of the ringing artefact might have been reduced more than the presence of the blocking artefact (see Figure 5.2) for each type of compressed/reference footage. Further, video compression algorithms embody deblocking filters (smoothing of sharp edges) and H.264/AVC features a deblocking filter on both the decoding path and on the encoding path [280]. This indicates that the artefacts have already been smoothed out from the compression algorithm itself. Also, there is some consistency in this investigation as the rescaling process is applied to both the 'uncompressed' reference and the compressed scenes. The compressed facial regions are accessed based on similarity score distance from their reference versions.

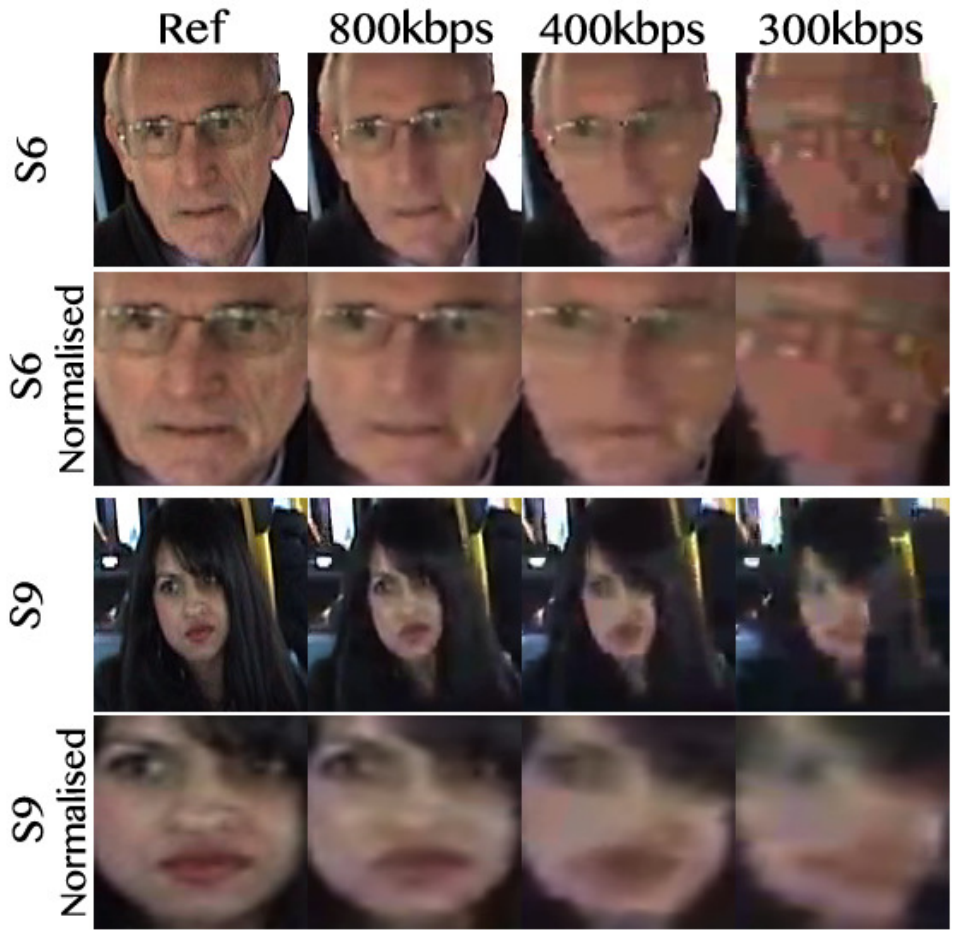


Figure 5.2: Rescaling of facial images. The rescaling of facial images has smoothed out edges and compression artefacts for both reference and compressed versions.

Initial tests have indicated that size and representation (i.e. grayscale images produced lower matching scores than colour images) of facial regions influence the derived matching scores. These were kept constant (see above paragraph) between reference and degraded footage with MPEG Streamclip encoder but not with the CCTV DVR encoders. More specifically, the eye coordinates were at different locations for the CCTV DVR encoders, which might have influenced the selection of the facial regions. For example, the selected facial regions (from each CCTV DVR encoder) are not exactly the same with the reference, degraded footage with MPEG Streamclip encoder, and the other CCTV DVR encoders. The CCTV DVR encoders are representative of the uncontrollable nature of CCTV systems in terms of not knowing what processes have been applied to the stored imagery (footage).

2. *Creation of the gallery dataset for testing the AFR systems.* A single dataset was created for the testing. The dataset was represented with a single folder containing 25 subfolders corresponding to each of the 25 scenes in Figure 4.4. The first 8 images for each scene/subfolder corresponded to the facial images in the reference format and were used to train the AFR systems (these are called enrols or known faces). The remaining images in the gallery dataset acted as the probes, or unknown faces. Furthermore, each subfolder consisted of 149 facial region images (as per the right image in Figure 5.1). More specifically, the 149 facial images in each subfolder were in the following order:

   a) 8 images from the reference (MPEG-2 format),

   b) 8 images from the reference (MPEG-2 format)-to identify how the systems behave when the reference/training 8 images are compared with themselves,

   c) 8 images from the original recorded "uncompressed" (DV format),

   d) 8 images from each of the 10 compressed types using the MPEG Streamclip implementation (at kbps: 300, 400, 600, 800, 100, 1200, 1400, 1600, 1800, 2000),

   e) and the remaining images were derived from the 5 CCTV recorders for the 6 identified most affected scenes (the key scenes). The remaining scenes

included repetitions of the reference (in order to create balancing number of images in the subfolders).

Inclusion of the images from the CCTV recorders did not affect the similarity scores. These were perceived, in the human investigation, to have a different quality from that of the MPEG Streamclip encoder. For example, the similarity scores were the same between the gallery dataset under investigation and a dataset that included instead of the images from the CCTV recorder (for the 6 most affected scenes) repetitions of the reference. The number and size of subfolders influence the derived matching scores, but not the order of subfolders, or the order of the images in the subfolders. It was considered important to include all the degraded footage (from MPEG Streamclip and CCTV recorders) in order to follow the same methodology/steps implemented in the human investigation

3. *Testing of the AFR systems.* There is a variety of testing procedures that can be used to evaluate automated systems [37, 38]. In this investigation, the entire aforementioned gallery dataset was processed by the AFR systems. Every single image that belonged to the known faces was compared against every single image of unknown faces and the produced similarity scores were presented in a matrix form. The distance/similarity between 2 face biometric signatures was calculated using the Euclidian distance (see Equation 5.1). The latter, in comparison to other distance measures (i.e. cosine mahalanobis, cosine, City block distance), provided more comparable values among the 3 AFR systems.

4. *Preparation of results for statistical analysis.* The derived similarity values were scaled in order to range between 0 (no similarity between known and unknown face images) and +1 (perfect similarity between known and unknown face images) using a simple normalisation formula (Eq. 5.2):

$$Normalisation = 1 - \frac{x_i - min(x)}{max(x) - x_i} \tag{5.2}$$

Where x_i, is a single similarity value in a sample x, and $min(x)$ and $max(x)$ correspond to the minimum and maximum values of that sample respectively. The sample consists of all the similarity values obtained from the entire testing of the gallery dataset for each AFR system. Figure 5.3, illustrates an example relating to the generated similarity matrices from the automated systems. As mentioned previously, each scene contained 8 consistent facial images (i.e. corresponding to 8 successive frames in the footage) of the same individual. In Figure 5.3, similarity scores for scene S1 are obtained between the 8 known/training facial images in reference format with the 8 unknown facial images in reference format, and with the 8 unknown facial images in H.264/MPEG-4 AVC at 300kbps format. The mean value of the derived similarity scores between each 2 set of 8 images (8 known×8 unknown) was used for further analysis. In case of the CCTV DVR encoders that will be the mean value between the 1 outputted image for each scene compared with the 8 images of the reference (8 known×1 unknown); since the CCTV DVR recorders provided only 1 facial image from the 8 reference facial images.

Scene 1

	Known 1	Known 2	Known 3	Known 4	Known 5	Known 6	Known 7	Known 8	
Unknown_Ref 1	1	0.9999	0.9998	0.9999	0.9999	0.9999	0.9998	0.9998	*Average*
Unknown_Ref 2	0.9999	1	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	*0.9999*
Unknown_Ref 3	0.9998	0.9999	1	0.9999	0.9999	0.9999	0.9999	0.9999	
Unknown_Ref 4	0.9999	0.9999	0.9999	1	0.9999	0.9999	0.9998	0.9998	
Unknown_Ref 5	0.9999	0.9999	0.9999	0.9999	1	0.9999	0.9999	0.9999	
Unknown_Ref 6	0.9999	0.9999	0.9999	0.9999	0.9999	1	0.9998	0.9999	
Unknown_Ref 7	0.9998	0.9998	0.9999	0.9998	0.9999	0.9998	1	0.9999	
Unknown_Ref 8	0.9998	0.9998	0.9999	0.9998	0.9999	0.9999	0.9999	1	
Unknown_300 1	0.9040	0.9040	0.904	0.904	0.904	0.904	0.904	0.904	Average
Unknown_300 2	0.9038	0.9038	0.9038	0.9038	0.9038	0.9038	0.9038	0.9038	*0.9073*
Unknown_300 3	0.9203	0.9203	0.9203	0.9203	0.9203	0.9203	0.9203	0.9203	
Unknown_300 4	0.9160	0.916	0.916	0.916	0.916	0.916	0.916	0.916	
Unknown_300 5	0.9052	0.9053	0.9052	0.9052	0.9052	0.9052	0.9052	0.9052	
Unknown_300 6	0.9006	0.9006	0.9005	0.9006	0.9006	0.9006	0.9006	0.9006	
Unknown_300 7	0.9094	0.9094	0.9094	0.9094	0.9094	0.9094	0.9094	0.9093	
Unknown_300 8	0.9116	0.9116	0.9115	0.9116	0.9115	0.9116	0.9115	0.9115	

Figure 5.3: Example of the generated similarity matrices. The values are the derived similarity scores obtained between the 8 known facial images from scene S1 (in reference format) with the 8 unknown facial images from scene S1 in the reference format (acting now as Unknown Ref) and in H.264/MPEG-4 AVC compressed at 300kbps format (Unknown 300). Each scene contains 8 consistent facial images. The average value is obtained by taking the mean of all the similarity scores between each 2 set of 8 images (8 known×8 unknown-highlighted by the grey box).

5.3 Results

In Sections 5.3.1 and 5.3.2, the results obtained from all the face recognition systems were modelled using the sigmoid logistic Eq. 4.1 in Section 4.3. Where matching scores (i.e. for AFR systems) and proportion of *yes* responses (i.e. for HFR system) are plotted vs. the different levels of compression. Eq. 4.1 seems to fit the data from the AFR systems well (this will become apparent from the created plots) and has contributed to keeping the statistical analysis consistent among all the face recognition systems. In non-linear modelling the choice of a function is based on how well a model fits the data [281]. The upper part of the human sigmoid logistic equation (Eq. 4.1) has fitted the data from the AFR systems well. Data modelling for the automated systems were curried out using R software for statistics [282]

by implementing non-linear least-squared regression [283]. The data obtained from the automated systems have a Gaussian distribution and a non-linear relationship was identified between compression rate and matching scores. The non-linear least-squared regression method fits models that minimise the sum of the squares of residuals. Errors on the fitted model coefficients are derived by an iterative procedure were the user supplies an initial guess; this is not the case for linear models as an initial guess for the coefficients is not required [284]. Common goodness of fit methods used by linear models such as R-Squared have been found inadequate for non-linear models [269]. For this reason a goodness of fit method has not been applied to the obtained non-linear models, instead the models with their associated errors on the coefficients have been utilised for the analysis/explanation of results.

Data modelling for the human results were processed accordingly to the psychometric curve fitting method described in Sections 4.3 and 4.4. The *Llambda* coefficient was set to range between 0.00 and 0.02 depending on how well the model fitted the raw data by taking into consideration the pDev value (see Section 4.3). Standard error calculations of the model coefficients for the AFR systems were calculated for all 3 coefficients (i.e. α, β, and Λ) but for the HFR system only for coefficients α and β (as Λ was given a set value).

In Section 5.3.3 the derived raw data from the 5 CCTV DVR systems with the key scenes are modelled using linear regression. The non-linear model of the sigmoid logistic Eq. 4.1 over-fitted the derived raw data and thus was concluded to be not appropriate. Instead linear regression fitted the data well. The aim of this step of analysis is to focus on observing the derived results and trends rather than identifying the best model for the data. For example, some of the outputted raw data have linear properties and some are close to linear. Overall the data obtained from the CCTV DVRs are scattered and linear models are fitted to observe tendencies. This will become apparent in Section 5.3.3.

The analysis of the results has been divided into 4 parts: 1) Section 5.3.1 includes

an analysis on the overall performance of each recognition system with respect to compression utilising the industry standard MPEG streamclip software (SC), 2) Section 5.3.2 involves a detailed analysis, based on the grouped properties in Table 4.3, on the performance of each recognition system with respect to compression utilising the industry standard MPEG streamclip software (SC), 3) Section 5.3.3 is concerned with the face recognition systems performance with the key scenes from the 5 CCTV DVR encoders, and 4) Section 5.3.4 provides an additional analysis relating to the performance of the AFR systems with the low lightness scenes.

5.3.1 Overall performance with industry standard H.264/MPEG-4 AVC encoder

Figure 5.4 presents results obtained from face recognition systems AFR - LDA, AFR - KFA, AFR - PCA, and HFR respectively. In all graphs, the raw data (matching scores for AFR and proportion of *yes* responses for HFR) of all the 25 scenes are plotted vs. the different levels of compression (in log kbps) and the reference (displayed separately and next to the main graphs). The lines in the graphs are the models obtained from modelling the raw data.

Table 5.1, includes details of the overall fitted models in Figure 5.4. The first column provides the system name. The second, fourth and sixth columns provide information on the derived coefficients of each model (α - absolute threshold, β - gradient and Λ - lapse rate). Their next columns provide the calculated standard error on the coefficients (std).

From the non-linear regression analysis of the AFR systems, there is a significant correlation (i.e. derived by calculating the statistical p value identifying any significant trends - see Table 5.1) between matching scores and compression rate for AFR systems LDA and KFA, but not for PCA. This is also visible from the graphs in Figure 5.4. The statistical tool utilised for the analysis of the human data does not include a calculation of the statistical p value. By observing the human model in

Figure 5.4 a conclusion can be made that a relationship exists between proportion of *yes* responses and compression rate. The derived human model does not provide a perfect fit but rather illustrates the tendency of the data, such as a sigmoid curve behaviour. The human data are too scattered to be able to obtain a perfect model in terms of being aligned to most of the data points but they do provide a systematic trend. This is the same for the PCA method as the obtained standard errors on the coefficients, in Table 5.1, are quite big. The obtained standard errors on the human model coefficients are not big, indicating that the model represents the overall data well.

This scattered behaviour to the human data is indicative to scene dependency that it is common in human evaluations; some scenes with specific properties are affected more than others by compression [48]. The scattered behaviour of PCA can be explained as being proof that this method does not perform well in minimising within-class variations of the same individual (see Section 2.1.2). LDA and KFA are designed to minimise within-class variations and have performed well when the reference was compared with itself. For example, both LDA and KFA label in the same class all the facial images that come from a single individual (i.e. in this case the 8 images of the same face in a single scene such as S1) and in a different class individuals that differ (i.e. face images from different scenes). This is not the case for PCA and assigns multiple images of the same individual to different classes. For this reason, even when the reference is compared with itself, PCA has performed the worst. This is because the mean value among the 8 images depicting the same individual was used. Yet, the scattered results of PCA still demonstrate a systematic trend (see Figure 5.4).

In the overall performance analysis, results show that the automated recognition systems are more tolerant to compression than humans (Figure 5.4). Performance drops at high compression rates (e.g. 300kbps, 400kbps, 600kbsp) for most systems. Each recognition system performed differently and KFA seems to have performed better than LDA.

The produced matching scores for all AFR systems, between reference MPEG-2 format and original recorded DV format were very similar, indicating no significant difference between the original recorded and the assigned reference footage (see Section 5.2).

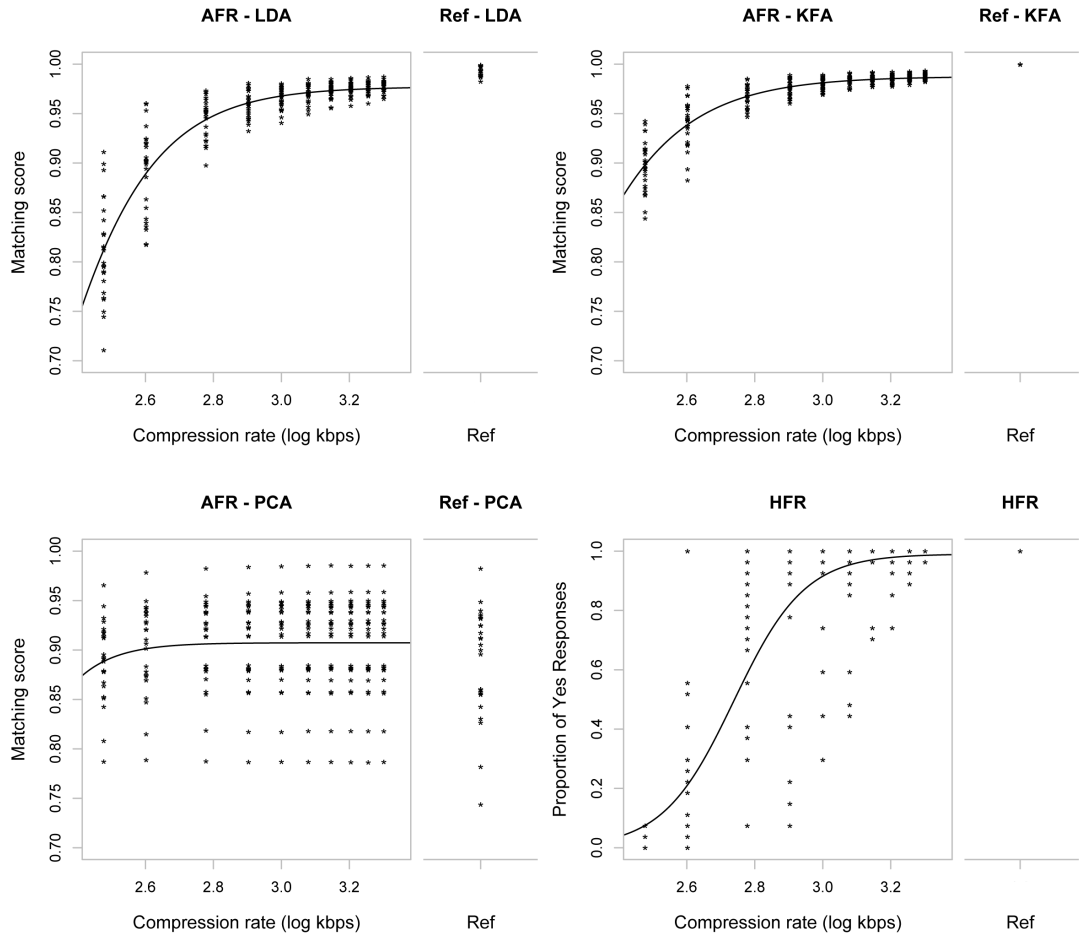


Figure 5.4: Overall performance of face recognition systems. The raw data (matching scores for AFR and proportion of *yes* responses for HFR) of all the 25 scenes are plotted with respect to the different levels of compression (in log kbps) and the reference (displayed separately and next to the main graphs). The lines in the graphs are models obtained from the raw data.

System	α	std	β	std	Λ	std
AFR - LDA	2.204	0.022	5.819	0.441	0.023	0.003
AFR - KFA	2.035	0.035	5.211	0.408	0.012	0.002
AFR - PCA	2.061	0.580	9.254	12.710	0.093	0.004
HFR	2.738	0.005	9.566	0.227	0.010	NA

Table 5.1: Coefficient information of the overall fitted models in Figure 5.4. The first column provides the system name, where α, β and Λ are the coefficients for the absolute threshold, gradient and lapse rate respectively. std columns provide the calculated standard error on the coefficients. Where NA stands for not applicable. The std of the Λ coefficient has not been calculated for the human data as it was kept constant (see Section 4.3).

5.3.2 Group category performance with industry standard H.264 (MPEG-4 AVC) encoder

This section includes a detailed analysis on the performance of each recognition system for each grouped scene property detailed in Table 4.3 (e.g. close and far properties of the camera to subject distance groups). This analysis will allow identification of correlations of image acceptance (or scene property acceptance) between automated and human face recognition systems. Additionally, it will identify the scene properties that decline performance of automated systems. This information can be utilised by designers of such systems in terms of identifying techniques to improve performance for the 'declined' scene properties.

Figures 5.5 to 5.7 include raw points and their fitted models as in Figure 5.4, but this time is based on individual scene properties. Tables 5.2 to 5.4 provide details on the calculated coefficients together with their associated calculated standard errors of the models in Figures 5.5 to 5.7. For example, in Figure 5.5 the behaviour of the recognition systems AFR - LDA and HFR under the angle of the face to camera plane groups with respect to compression rate is investigated. The raw data (y-axis: matching scores for AFR and proportion of *yes* responses for HFR) of each scene property under the angle group is plotted with respect to compression rate (x- axis: in log kbps.)

When the calculated coefficients (+/- std) between models overlap (especially for

α and β coefficients) then this is an indication of there being no difference between the investigated scene properties or produced models. For example, in Table 5.2 and system AFR - KFA the coefficients for 'tilted' ($\alpha = 2.033 + / - 0.055$, $\beta = 5.001 + / - 0.613$ and $\Lambda = 0.013 + / - 0.003$) and 'frontal' ($\alpha = 2.039 + / - 0.040$, $\beta = 5.451 + / - 0.489$ and $\Lambda = 0.012 + / - 0.002$) angle groups are not different thus these 2 models are classified as being the same. For the angle property, only the fitted models of system AFR - LDA do not overlap (see Figure 5.5). The AFR - LDA 'frontal' scenes have produced higher scores than 'tilted' scenes (i.e. 'tilted' angle scenes are affected more by compression than 'frontal' scenes). Similarly, for the distance property only the fitted models of system HFR do not overlap (see Figure 5.5). The HFR 'close' scenes have produced higher scores than the 'far' distance scenes (i.e. 'far' distance scenes are affected more by compression than 'close' scenes).

System	α	**std**	β	**std**	Λ	**std**
$LDA_{Frontal}$	2.236	0.027	6.655	0.684	0.023	0.003
LDA_{Tilted}	2.169	0.035	5.107	0.562	0.022	0.004
LDA_{Close}	2.217	0.030	5.987	0.650	0.019	0.004
LDA_{Far}	2.190	0.032	5.654	0.588	0.0263	0.004
$KFA_{Frontal}$	2.039	0.040	5.451	0.489	0.012	0.002
KFA_{Tilted}	2.033	0.055	5.001	0.613	0.013	0.003
KFA_{Close}	2.041	0.049	5.237	0.583	0.010	0.002
KFA_{Far}	2.028	0.049	5.186	0.563	0.015	0.002
$PCA_{Frontal}$	2.101	0.754	10.308	20.372	0.086	0.005
PCA_{Tilted}	2.022	0.881	8.391	16.007	0.099	0.005
PCA_{Close}	2.144	0.608	11.675	21.078	0.084	0.004
PCA_{Far}	1.977	1.023	7.624	15.374	0.100	0.006
$HFR_{Frontal}$	2.762	0.007	7.845	0.248	0.006	NA
HFR_{Tilted}	2.721	0.060	11.930	0.410	0.006	NA
HFR_{Close}	2.702	0.007	10.461	0.437	0.012	NA
HFR_{Far}	2.773	0.006	9.161	0.273	0.006	NA

Table 5.2: Coefficient information of the fitted models for Distance and Angle category groups. The same approach is adopted as was used in Table 5.1

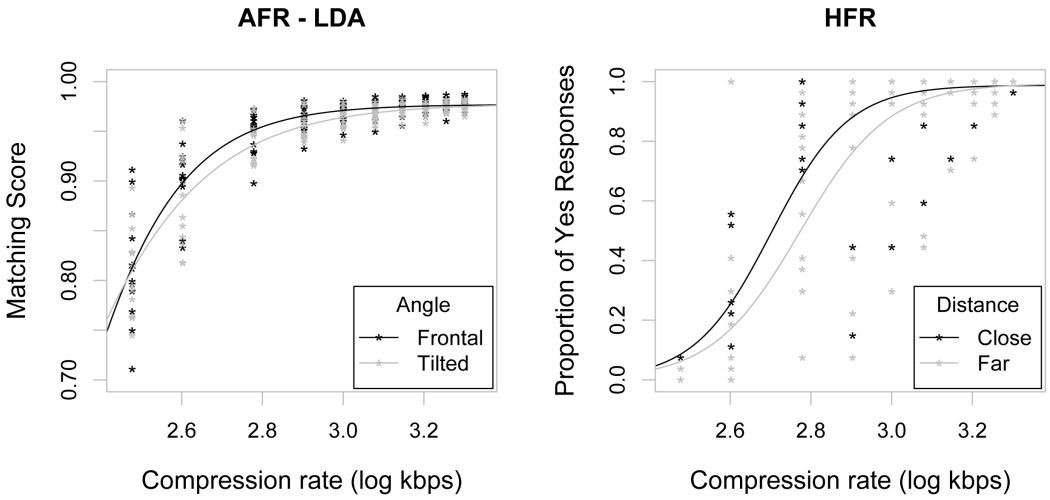


Figure 5.5: Angle of face to camera plane and camera to subject distance groups. For all graphs, the raw data points (x- axis: matching scores for AFR and proportion of *yes* responses for HFR) are plotted with respect to compression rate (y- axis: in log kbps). The black points (*) and lines represent raw data and models respectively of scenes with the tilted angle and close distance property. The grey points (*) and lines represent raw data and regression models respectively of scenes with the tilted angle and far distance property.

For the busyness groups (Figure 5.6 and Table 5.3), the AFR - LDA and AFR - KFA systems have produced different curve models for properties 'High spatial-High temporal' and 'Low spatial-Low temporal' (curve models for 'Low spatial-High temporal' and 'High spatial-Low temporal' are the same). 'Low spatial-Low temporal' scenes have produced the highest scores and can afford more compression than the rest of the busyness groups. A similar behaviour can be observed for system AFR - PCA in terms of the 'Low spatial-Low temporal' scenes even though the error estimates on the β coefficients are quite big.

For the HFR system, the 'High temporal-High spatial' model is different from the rest models and the rest models are all overlapping with each other (either threshold - α or gradient - β or both). The 'High temporal-High spatial' scenes are influenced more by compression than the remaining busyness category groups.

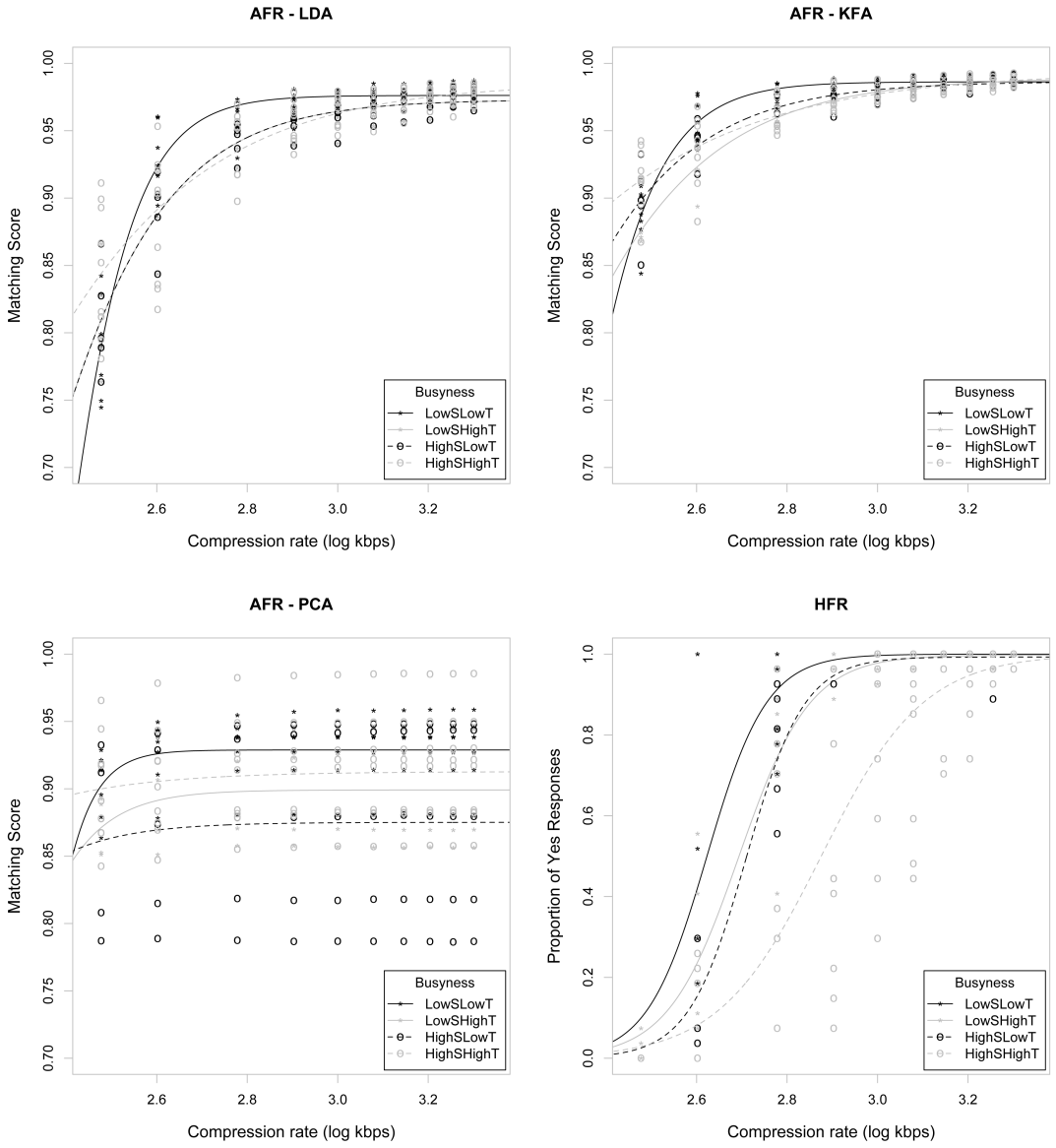


Figure 5.6: Busyness groups. For all graphs, the raw data points and their fitted models are plotted with respect to compression rate. The black points (*) and solid lines represent raw data and models of scenes with the 'Low spatial-Low temporal' busyness property. The grey points (*) and solid lines represent scenes with the 'Low spatial-High temporal' busyness property. The black points (o) and dashed lines represent scenes with the 'High spatial-Low temporal' busyness property. The grey points (o) and dashed lines represent scenes with the 'High spatial-High temporal' busyness property.

System	α	**std**	β	**std**	Λ	**std**
$LDA_{LowSLowT}$	2.350	0.013	11.372	1.034	0.024	0.002
$LDA_{LowSHighT}$	2.224	0.030	5.071	0.561	0.019	0.006
$LDA_{HighSLowT}$	2.208	0.035	5.935	0.721	0.027	0.004
$LDA_{HighSHighT}$	1.999	0.084	3.729	0.682	0.014	0.008
$KFA_{LowSLowT}$	2.262	0.0217	10.169	0.986	0.0137	0.0013
$KFA_{LowSHighT}$	2.050	0.041	4.781	0.454	0.011	0.003
$KFA_{HighSLowT}$	2.037	0.072	5.275	0.845	0.014	0.003
$KFA_{HighSHighT}$	1.711	0.139	3.175	0.614	0.006	0.005
$PCA_{LowSLowT}$	2.282	0.292	18.117	27.138	0.071	0.003
$PCA_{LowSHighT}$	2.148	0.642	10.441	20.051	0.101	0.007
$PCA_{HighSLowT}$	1.895	3.494	7.114	42.222	0.125	0.013
$PCA_{HighSHighT}$	1.539	2.798	4.533	13.644	0.087	0.009
$HFR_{LowSLowT}$	2.623	0.007	15.116	0.988	0.000	NA
$HFR_{LowSHighT}$	2.693	0.010	12.732	0.905	0.002	NA
$HFR_{HighSLowT}$	2.710	0.009	15.640	1.092	0.007	NA
$HFR_{HighSHighT}$	2.876	0.007	8.828	0.303	0.000	NA

Table 5.3: Coefficient information of the fitted models for the busyness category groups. The same approach is adopted as was used in Table 5.1

As the number of properties under each grouped category increases, it becomes more difficult to derive conclusions on the derived models. For example, the groups in the lightness category consists of 5 properties (e.g. bus, medium, low, high and mixed lightness) distributed across 25 scenes and most of the derived models overlap (Figure 5.7 and Table 5.4). In a future investigation more scenes should be included. Nonetheless, the graphs in Figure 5.7 can still be observed to understand tendencies. For the automated systems, 'mixed lightness' scenes were the most affected (i.e. produced the lowest matching scores) and 'low lightness' scenes were the least affected (i.e. produced the higher matching scores) by compression. In contrast for humans, 'low lightness' scenes were the most affected and, 'medium' and 'mixed lightness' scenes the least affected by compression. Also, for the HFR system, the 'low lightness' model is different from the other lightness models and these other models are all overlapping with each other.

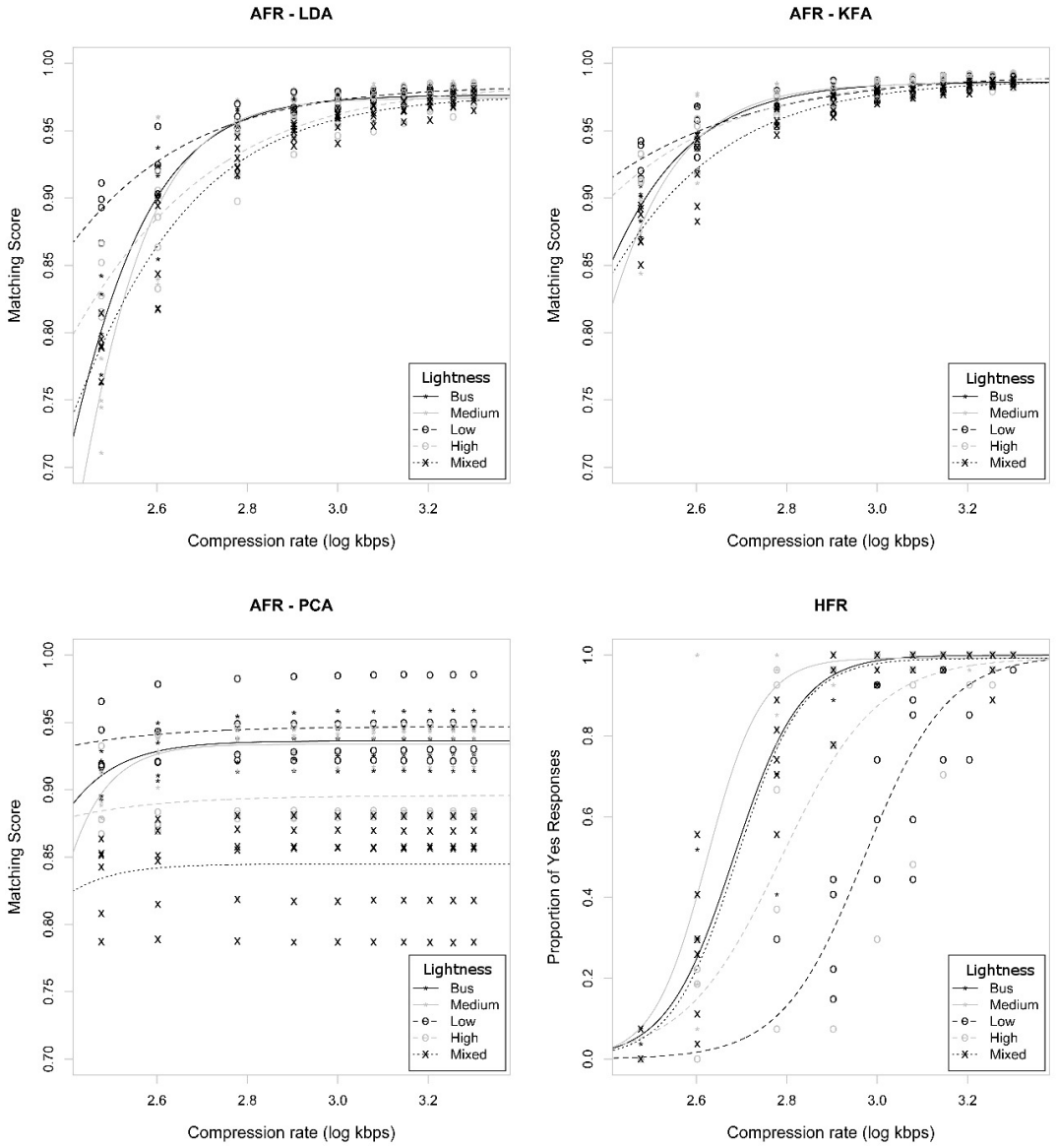


Figure 5.7: Lightness groups. For all graphs, the raw data points and fitted models are plotted with respect to compression rate. The black points (*) and solid lines represent raw data and models of scenes with the 'Bus medium lightness' property (Bus). The grey points (*) and solid lines represent raw data and models of scenes with the 'Daylight medium lightness' property (Medium). The black points (o) and dashed lines represent raw data and models of scenes with 'Daylight low lightness' property (Low). The grey points (o) and dashed lines represent raw data and models of scenes with the 'Daylight high lightness' property (High). The black points (x) and dotted lines represent raw data and models of scenes with 'Daylight mixed lightness' property (Mixed).

System	α	std	β	std	Λ	std
LDA_{Bus}	2.279	0.025	7.700	0.870	0.024	0.003
LDA_{Medium}	2.340	0.020	9.174	1.174	0.026	0.004
LDA_{Low}	1.944	0.081	4.263	0.655	0.016	0.004
LDA_{High}	2.041	0.067	3.893	0.609	0.015	0.007
LDA_{Mixed}	2.180	0.028	4.855	0.433	0.023	0.004
KFA_{Bus}	2.132	0.039	6.593	0.729	0.013	0.002
KFA_{Medium}	2.213	0.040	8.006	1.159	0.015	0.002
KFA_{Low}	1.665	0.187	3.308	0.804	0.008	0.005
KFA_{High}	1.821	0.100	3.895	0.610	0.009	0.003
KFA_{Mixed}	2.033	0.050	4.640	0.520	0.012	0.003
PCA_{Bus}	2.116	0.296	9.862	7.947	0.064	0.003
PCA_{Medium}	2.237	0.113	13.290	6.198	0.066	0.002
PCA_{Low}	1.546	3.222	4.813	16.741	0.053	0.008
PCA_{High}	1.603	2.701	4.969	15.395	0.104	0.008
PCA_{Mixed}	2.055	1.566	10.321	37.945	0.155	0.005
HFR_{Bus}	2.683	0.009	13.457	0.779	0.002	NA
HFR_{Medium}	2.624	0.009	16.780	2.003	0.007	NA
HFR_{Low}	2.972	0.009	10.968	0.619	0.000	NA
HFR_{High}	2.787	0.010	9.333	0.463	0.006	NA
HFR_{Mixed}	2.690	0.008	13.936	0.916	0.008	NA

Table 5.4: Coefficient information of the fitted models for the lightness category groups. The same approach is adopted as was used in Table 5.1

The detailed results from the HFR system in this current investigation agree with the results in Chapter 4 (see Table 4.8). Even though a different statistical approach has been implemented, the same conclusions are drawn. HFR system performance when assessed with compression is affected the most negatively by scenes exhibiting 'low lightness', 'far camera to subject distance' and 'High spatial-High temporal busyness' properties.

5.3.3 Key scenes performance with standard and CCTV DVR H.264 / MPEG-4 AVC encoders

This section provides the results from the AFR systems performance assessed with the key scenes and the 5 CCTV DVR H.264/MPEG-4 AVC encoders. Refer to Section 4.4 for a detailed description on the CCTV DVRs outputted compressed scenes (footage). Figures 5.8 to 5.10 illustrate the raw data points and fitted models plotted with respect to compression rate (in kbps). A single model was fitted to

all the data points derived from the 5 CCTV DVR (DVR) encoders for each key scene. Tables 5.5 to 5.7 provide information on the derived coefficients from the fitted models in Figures 5.8 to 5.10.

Figures 5.8 to 5.10 when compared with the results in Figure 5.4 illustrate that the automated systems have performed better with the industry standard (SC) encoder than the CCTV DVR encoders. This is applicable for the HFR system except for scene S12, where the DVR encoders have performed better. A more detailed analysis of the HFR system performance with footage from the DVR encoders can be found in Sections 4.4 and 4.5 and Table 5.8. For the AFR systems, as it has been mentioned previously, the selected facial region between the reference and compressed versions with the CCTV DVRs were not the same and this might have affected the derived data more than the compression applied and reduction of frame rate (i.e. to 4fps). A conclusion cannot be made as the uncontrollable nature of CCTV DVRs introduces hidden variables.

All the automated systems have performed the best with the 'low lightness' scenes S12 and S13 and the worst with the 'mixed lightness' scene S25. It appears difficult to derive the same conclusions for the HFR system by observing the already modelled results in Sections 4.3 and 4.4. Instead, Table 5.8 provides a comparison of SC and DVR encoders at 60% and 75% proportions of *yes* responses for each key scene. The HFR system has performed the best with the 'medium lightness' scenes (Bus and Daylight illumination) and the worst with the 'low lightness' scenes S12 and S13 together with the 'high lightness' scene S17 for both encoders at 60% proportion of *yes* responses and at 75% proportion of *yes* responses for SC encoder. Whereas, for the CCTV DVR encoder at 75% proportion of *yes* responses scenes S13, S17 and S25 were the most affected.

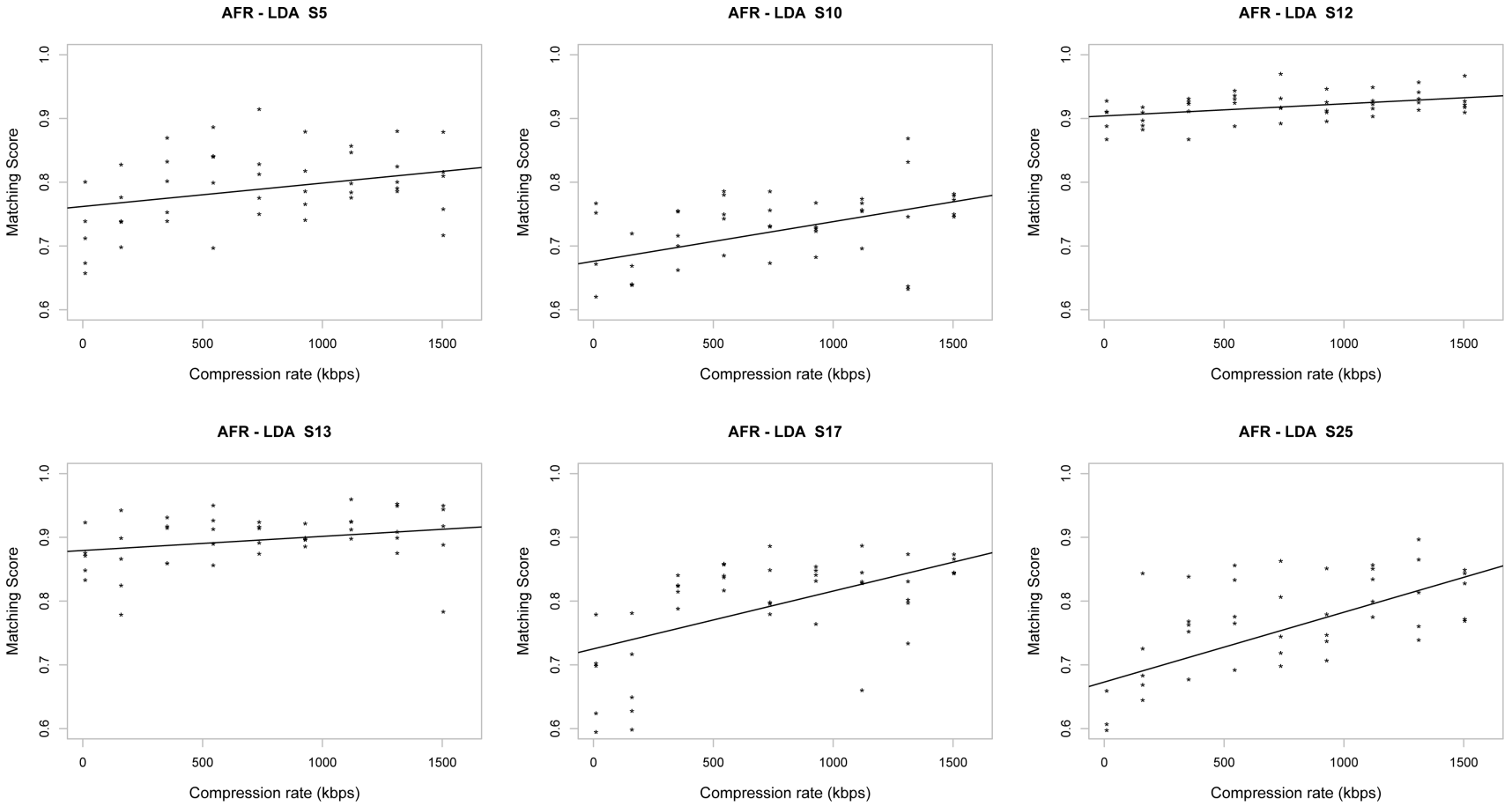


Figure 5.8: AFR-LDA performance with key scenes. For all graphs, the raw data points (*) and fitted models are plotted with respect to compression rate (in kbps). A single linear model was fitted to all the data points derived from the five CCTV DVR (DVR) encoders for each key scene.

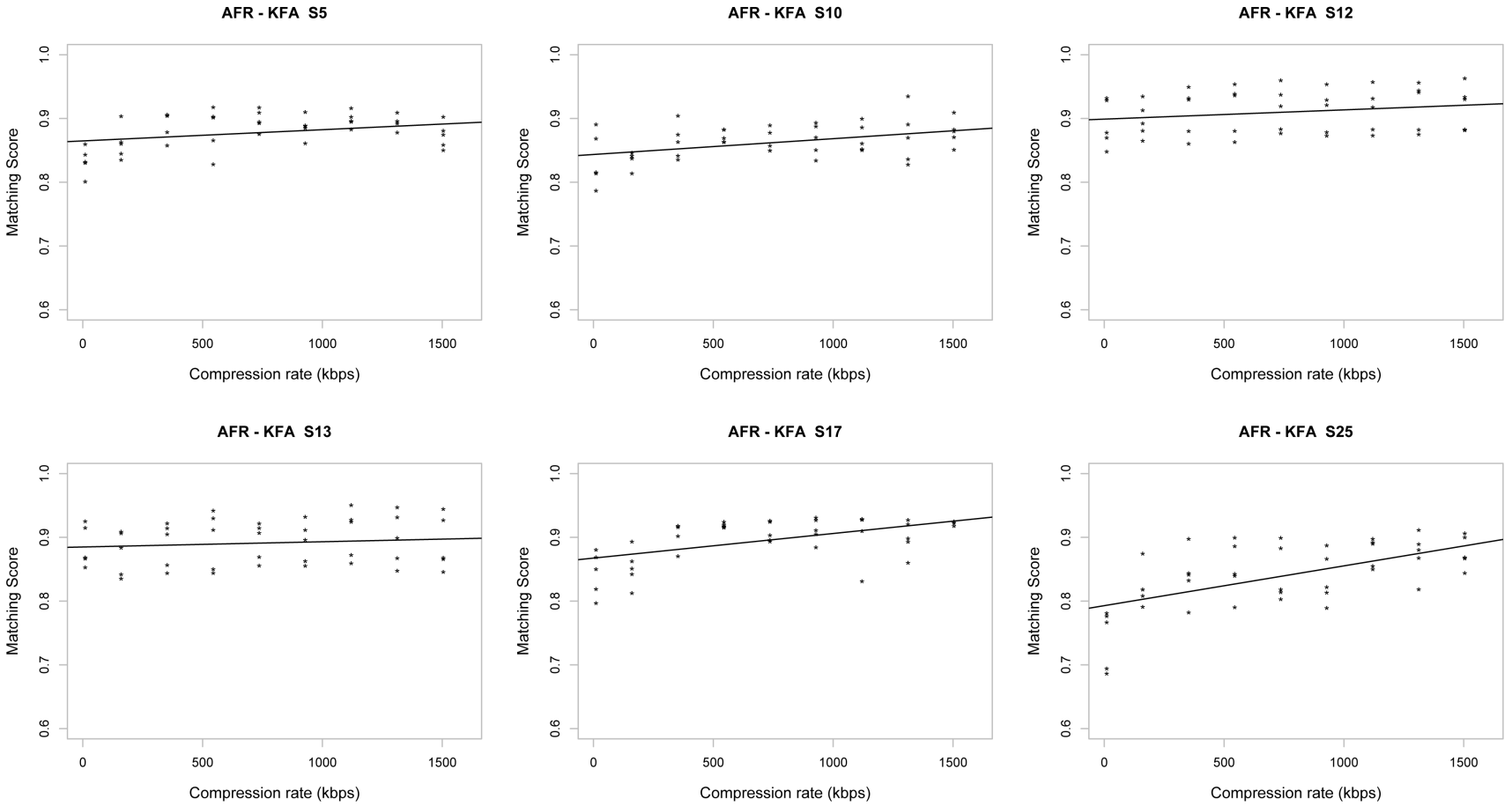


Figure 5.9: KFA-LDA performance with key scenes. Adopting the same approach as in Figure 5.8.

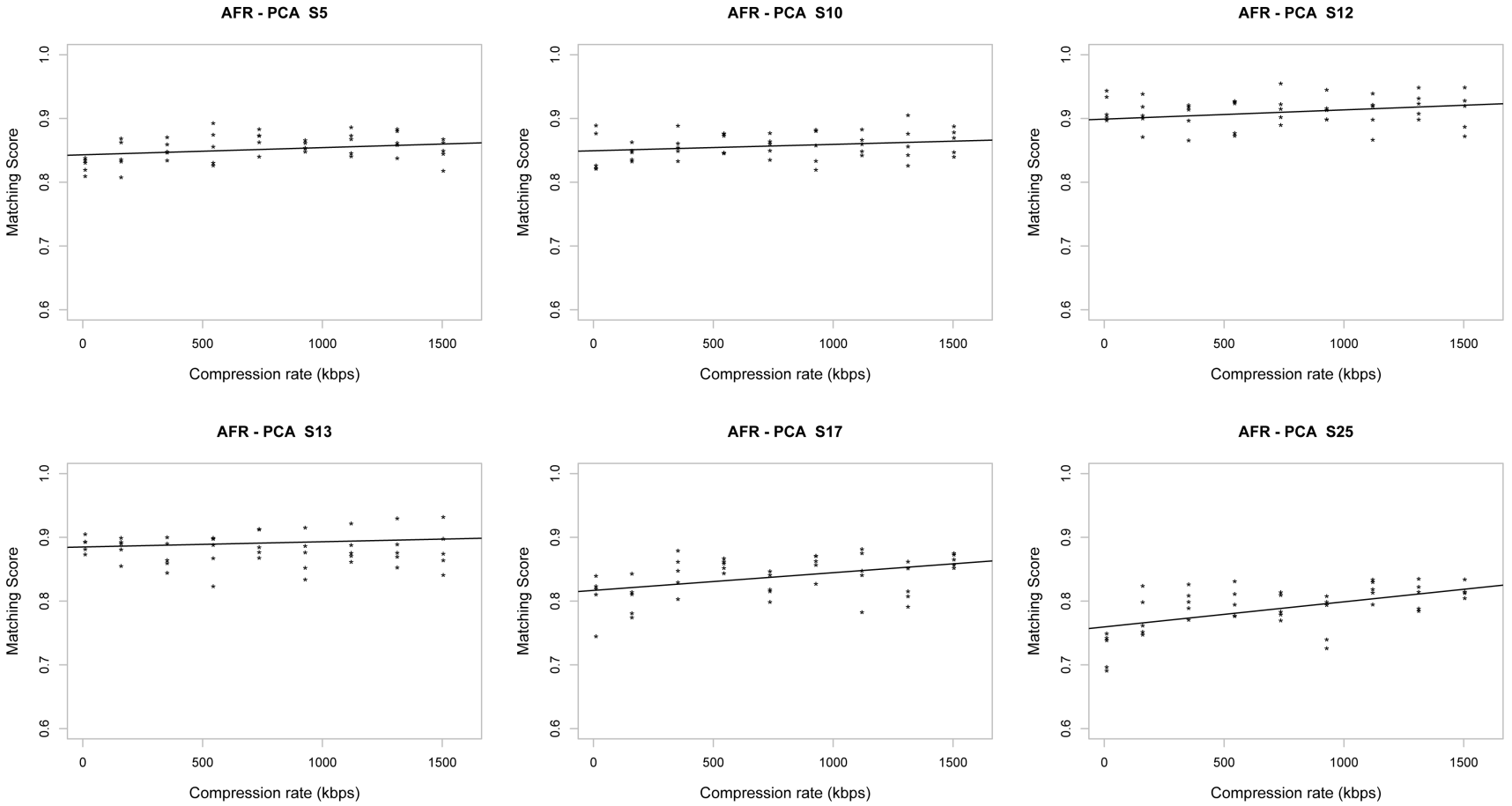


Figure 5.10: AFR-PCA performance with key scenes. Adopting the same approach as in Figure 5.8.

System	α	**std**	β	**std**	p
S5 - LDA_{DVR}	0.760	0.015	0.000	0.000	0.0212*
S10 - LDA_{DVR}	0.671	0.015	0.000	0.000	0.0000****
S12 - LDA_{DVR}	0.904	0.005	0.000	0.000	0.001**
S13 - LDA_{DVR}	0.879	0.012	0.000	0.000	0.062.
S17 - LDA_{DVR}	0.725	0.017	0.000	0.000	0.000***
S25 - LDA_{DVR}	0.673	0.019	0.000	0.000	0.000***

Table 5.5: Coefficient information of the fitted models for key scenes from system AFR-LDA. Where α and β are the obtained coefficients of the fitted linear models. std stands for standard error for each of the calculated coefficients. Where p is the statistical value and defines whether a significant trend exists between matching scores and compression rate.

System	**a**	**std**	**b**	**std**	p
S5 - KFA_{DVR}	0.865	0.007	0.000	0.000	0.017*
S10 - KFA_{DVR}	0.844	0.007	0.000	0.000	0.001***
S12 - KFA_{DVR}	0.899	0.009	0.000	0.000	0.114
S13 - KFA_{DVR}	0.885	0.009	0.000	0.000	0.399
S17 - KFA_{DVR}	0.867	0.007	0.000	0.000	0.000***
S25 - KFA_{DVR}	0.793	0.011	0.000	0.000	0.000***

Table 5.6: Coefficient information of the fitted models for key scenes from system AFR-KFA. Adopting the same approach as in Table 5.5.

System	**a**	**std**	**b**	**std**	p
S5 - PCA_{DVR}	0.084	0.005	0.000	0.000	0.043*
S10 - PCA_{DVR}	0.849	0.005	0.000	0.000	0.084
S12 - PCA_{DVR}	0.909	0.006	0.000	0.000	0.627
S13 - PCA_{DVR}	0.882	0.007	0.000	0.000	0.959
S17 - PCA_{DVR}	0.817	0.007	0.000	0.000	0.001***
S25 - PCA_{DVR}	0.759	0.007	0.000	0.000	0.000***

Table 5.7: Coefficient information of the fitted models for key scenes from system AFR-PCA. Adopting the same approach as in Table 5.5.

			60% (kbps)			
Encoder	S5	S10	S12	S13	S17	S25
SC	670	752	1255	1231	1285	711
CCTV DVR	480	596	714	1131	969	657

			75% (kbps)			
Encoder	S5	S10	S12	S13	S17	S25
SC	753	811	1467	1469	1437	788
CCTV DVR	751	771	901	1674	1409	1045

Table 5.8: Comparison between industry and CCTV DVR encoders at 60% and 70% points of *yes* responses for each key scene

5.3.4 Additional analysis

Similarly to the findings of Adler and Dembinsky [44], the results in Sections 5.3.2 and 5.3.3 illustrate that the performance of AFR and HFR systems differ and it is dependent on scene content. 'Low lightness' scenes have been affected the least by compression for AFR systems and the most for the HFR system. On the other hand, the HFR system performed the best with 'medium lightness' (including both Bus and Daylight illumination) scenes.

Dietz and Eberhart [285] investigated the impact of a camera's different ISO values on image quality. They have found that processing of extremely under exposed captures (due to a setting of low ISO) can produce comparable image quality to those with the camera ISO increased for a correct exposure keeping constant the shutter speed and aperture settings. It is well known among photographers that under exposed scenes entail more visual information than over exposed scenes.

Clipping in photography is when areas in an image (or entire image) appear uniformly with minimum and/or maximum levels of brightness (e.g. blue sky appearing white because of over exposure). Clipping can occur due to an incorrect capture (i.e. over and under exposed scenes) or a digital process (e.g. sharpening). The degree by which values are clipped affects the amount of information in an image is lost/hidden. Another factor that affects the degree of clipping is the limitations

of the colour gamut properties of the imaging system. This relates to the capability of the imaging system to be able to distinguishing colours/shades at different brightness levels. For example, display outputs use a smaller colour gamut than an average digital camera. Perhaps the information in an under exposed scene is within the image but not reproducible/visible on a display because of smaller gamut [286, 287]. Furthermore, other factors such as errors on the quantisation process could cause blocks of uniform brightness. This is known as quantisation noise and could be caused during the digitisation process of an analogue signal to the output digitised values.

Figure 5.11 and 5.12 illustrate examples of applying a simple processing (by altering the image levels) method on 'low lightness' key scenes S12 and S13. The processing method has revealed facial information that was not visible in the original sequence of facial images. The facial information in the processed images can be compared with the 'medium lightness' scenes S5 and S10 in Figure 5.13.

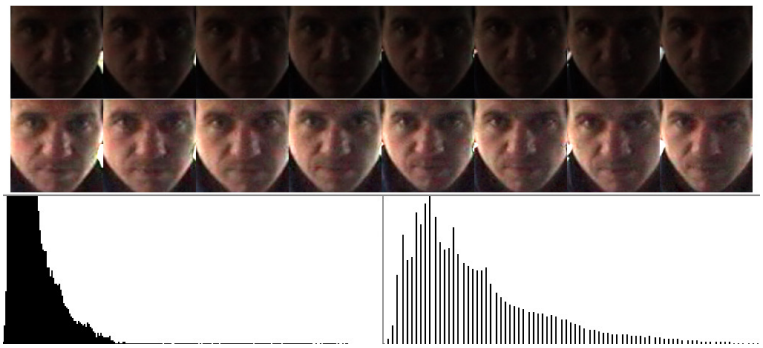


Figure 5.11: Processing of 'low lightness' scene S12. The top row of facial images represent the original sequence of the 8 facial images and the bottom row the same facial images after processing. The histogram on the left depicts the original facial images and the one on the right after processing.

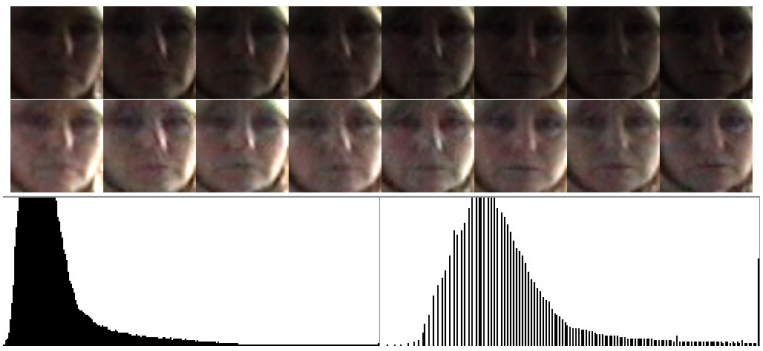


Figure 5.12: Processing of 'low lightness' scene S13. The top row of facial images represent the original sequence of the 8 facial images and the bottom row the same facial images after processing. The histogram on the left depicts the original facial images and the one on the right after processing.

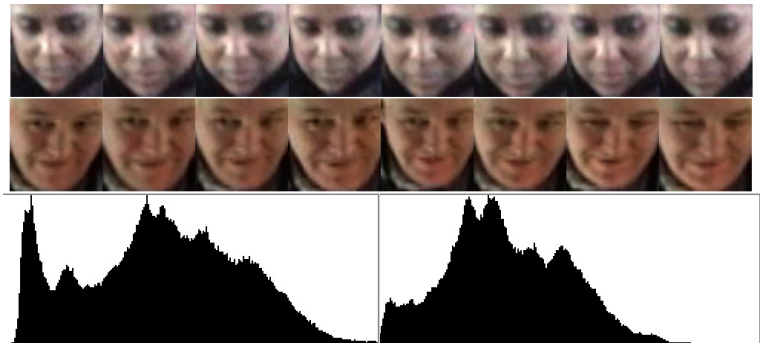


Figure 5.13: 'Medium lightness' scenes S5 (Bus illumination-top row) and S10 (Daylight-bottom row) together with their histograms. The top row of facial images represent the original sequence of the 8 facial images for scene S5 and the bottom row for scene S10. The histogram on the left depicts scene S5 and the one on the right scene S10.

Furthermore, Figure 5.14 presents the tone characteristics or transfer function, in a log-log scale, of the reference input to the automated systems. This tone transfer function represent the tone characteristics of the reference in either DV or MPEG-2 format and compressed versions of the reference with the industry standard encoder at 300kbps (i.e. not the CCTV DVRs). This means that the industry standard compressor (MPEG Streaclip) has not degraded the tone reproduction with respect to the reference 'uncompressed' footage. The γ of the characteristic curve is less than 1 ($\gamma = 0.8894$). A less than 1 γ indicates that the blacks or low luminance tones are expanded or quantised to a greater degree than high luminance tones. The above paragraph on under exposed scenes entailing more information after processing is part of the same argument. Also, the displayed tone characteristics

(i.e. input to the human visual system - see Figure A.2 with a $\gamma = 2.09$) are different from the measured tone characteristics (i.e. input to the automated systems - see Figure 5.14 with a $\gamma = 0.8894$).

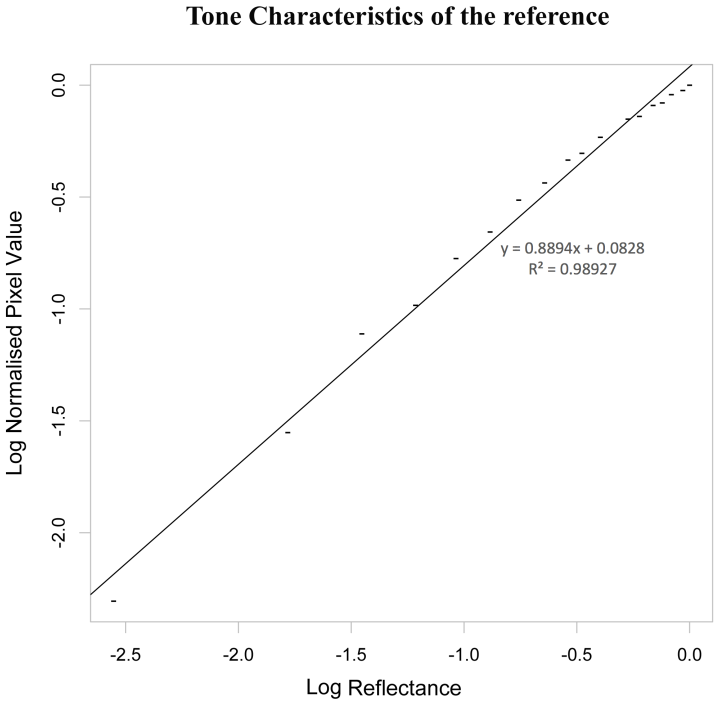


Figure 5.14: Tone characteristics of the reference

It is anecdotal why the under exposed scenes (e.g. S12 and S13) have scored higher than the correctly exposed scenes (e.g. S5 and S10) for the AFR systems. Perhaps, compression algorithms first remove the dark effect caused from under exposure and later proceed with scene content information. Compression tends to throw away what humans can not see without changing the tone reproduction. As it has been mentioned previously, compression has not affected the tone reproduction. Yet, even at high compression levels (300kbps and 400kbps) the 'low lightness' scenes have been affected the least for the AFR systems.

In order to understand further the behaviour of the AFR systems with scene content properties, a more detailed analysis has been considered. Tables 5.9 to 5.11 present the top 6 matching scores (or rank order results) for each reference version of the 25 scenes with the rest of the 25 reference scenes. No compressed versions are included in this analysis. The highest matching score for each of the 25 scenes is with itself.

Rank order identifies the top matches of an 'unknown' facial image. Observing the results from PCA (Table 5.9), LDA (Table 5.10), and KFA (Table 5.11) one can see that the obtained 6 top rank facial images relate to the lightness properties of the scenes/faces. Furthermore, PCA and KFA have obtained the top 6 rank matches, in most cases within the same lightness classification of the 'unknown' scene. For example, for both PCA and KFA the top 6 matches for scene S1 (i.e. exhibits 'medium lightness' property) are of the same 'medium lightness' category with either daylight or bus illumination. LDA has included for all the 'unknown' scenes the 'low lightness' scenes (S11, S12, S13, and S14) in the top 6 matches (see Table 5.10).

The lightness properties of faces/scenes have influenced the matching scores notably. In case of the 'low lightness' scenes for all the AFR systems, their top 6 matches among other 'low lightness' scenes are the highest, in comparison to the other lightness categories; such as 'medium lightness', where their top 6 matches include other 'medium lightness' scenes (see Tables 5.9, 5.10, and 5.11). For instance, Table 5.10 includes the top 6 matches for the LDA method and for the 'low lightness' scenes (S11,S12,S13,S14) the second top match value is around 0.7, whereas for the rest of the lightness scenes the second top match value is around 0.4. As tone characteristics do not change in the compressed scenes, perhaps the automated algorithms perform pattern/lightness matching between the dark areas (the employed AFR techniques are holistic after all) of the facial images in the 'low lightness' category, even at high compression levels. This might be able to be corrected by applying illumination normalisation techniques before automated face recognition. Illumination normalisation techniques could be used to compensate for the variable exposed footage that has been produced by incorporating in the methodology auto-white balancing and auto-exposure settings.

The issues in face recognition because of illumination variations are well known and still remain unsolved [189]. There is a huge amount of research investment in illumination normalisation techniques for optimising facial images for AFR systems;

histogram equalisation is the most commonly used technique [188–193]. Research has shown that normalisation of facial images from illumination variations does improve performance of face recognition systems. However, it is unknown how face normalisation illumination techniques influence performance based on specific scene content properties.

AFR-PCA: Rank Order of Matching Scores							
S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
S1:0.945	S2:0.913	S3:0.958	S4:0.937	S5:0.926	S6:0.943	S7:0.937	S8:0.926
S2:0.828	S1:0.828	S1:0.805	S13:0.784	S6:0.782	S10:0.841	S6:0.830	S9:0.842
S3:0.805	S3:0.760	S9:0.784	S10:0.778	S7:0.733	S7:0.830	S10:0.789	S10:0.776
S6:0.755	S6:0.757	S10:0.767	S8:0.764	S10:0.758	S9:0.818	S9:0.774	S4:0.764
S10:0.748	S10:0.755	S8:0.763	S9:0.758	S9:0.722	S5:0.782	S5:0.733	S6:0.764
S9:0.734	S9:0.729	S2:0.760	S14:0.747	S4:0.709	S8:0.763	S8:0.719	S3:0.763
S9	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**
S9:0.943	S10:0.916	S11:0.921	S12:0.950	S13:0.931	S14:0.985	S15:0.946	S16:0.880
S8:0.841	S6:0.841	S12:0.886	S14:0.923	S11:0.872	S12:0.923	S17:0.754	S17:0.721
S6:0.819	S9:0.808	S14:0.883	S11:0.886	S14:0.853	S11:0.883	S16:0.708	S18:0.719
S10:0.808	S7:0.790	S13:0.872	S13:0.851	S12:0.852	S13:0.853	S19:0.702	S15:0.718
S3:0.784	S4:0.778	S4:0.743	S3:0.727	S4:0.784	S4:0.747	S18:0.654	S19:0.711
S7:0.775	S8:0.776	S3:0.715	S4:0.725	S8:0.716	S8:0.736	S20:0.620	S21:0.598
S17	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
S17:0.880	S18:0.877	S19:0.874	S20:0.867	S21:0.853	S22:0.856	S23:0.875	S24:0.781
S15:0.756	S16:0.724	S15:0.720	S21:0.677	S20:0.680	S20:0.625	S16:0.691	S17:0.640
S16:0.716	S19:0.698	S16:0.719	S7:0.673	S23:0.632	S2:0.609	S18:0.680	S21:0.611
S19:0.696	S17:0.695	S17:0.706	S6:0.660	S6:0.619	S21:0.603	S15:0.674	S15:0.598
S18:0.689	S23:0.687	S18:0.700	S5:0.630	S17:0.618	S6:0.576	S17:0.667	S20:0.595
S23:0.664	S15:0.662	S23:0.630	S22:0.621	S7:0.612	S5:0.576	S21:0.639	S23:0.581
S25							
S25:0.813							
S7:0.601							
S5:0.580							
S23:0.567							
S6:0.566							
S4:0.546							

Table 5.9: AFR-PCA: Rank order of matching scores. The top 6 scene matches and their matching scores are provided for each of the 25 scenes.

AFR-LDA: Rank Order of Matching Scores							
S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
S1:0.985	S2:0.986	S3:0.985	S4:0.978	S5:0.973	S6:0.985	S7:0.991	S8:0.980
S12:0.442	S12:0.452	S12:0.443	S11:0.491	S14:0.420	S12:0.468	S11:0.380	S14:0.409
S11:0.436	S11:0.418	S11:0.418	S14:0.464	S12:0.391	S14:0.451	S12:0.378	S12:0.364
S14:0.431	S14:0.418	S14:0.416	S12:0.450	S11:0.385	S11:0.445	S10:0.359	S13:0.355
S13:0.409	S1:0.378	S13:0.405	S13:0.422	S2:0.350	S13:0.420	S13:0.342	S11:0.352
S3:0.393	S13:0.374	S1:0.392	S6:0.414	S1:0.330	S4:0.414	S14:0.337	S9:0.317
S9	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**
S9:0.977	S10:0.979	S11:0.984	S12:0.983	S13:0.979	S14:0.980	S15:0.984	S16:0.946
S12:0.457	S12:0.507	S12:0.745	S11:0.746	S14:0.748	S13:0.749	S11:0.305	S5:0.328
S11:0.448	S11:0.496	S13:0.735	S14:0.726	S11:0.735	S11:0.723	S10:0.291	S14:0.317
S13:0.388	S14:0.408	S14:0.721	S13:0.718	S12:0.719	S12:0.730	S12:0.290	S17:0.306
S14:0.379	S6:0.390	S10:0.4908	S10:0.511	S4:0.421	S4:0.463	S13:0.265	S12:0.305
S4:0.375	S13:0.384	S4:0.491	S6:0.466	S6:0.419	S6:0.450	S3:0.266	S11:0.296
S17	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
S17:0.972	S18:0.961	S19:0.935	S20:0.980	S21:0.968	S22:0.969	S23:0.953	S24:0.951
S12:0.405	S12:0.300	S4:0.337	S3:0.269	S12:0.283	S14:0.290	S12:0.291	S11:0.312
S11:0.396	S17:0.300	S10:0.308	S10:0.257	S11:0.272	S12:0.268	S13:0.286	S14:0.303
S14:0.378	S6:0.298	S18:0.295	S11:0.246	S14:0.269	S13:0.249	S11:0.285	S12:0.294
S13:0.373	S19:0.285	S15:0.294	S12:0.244	S6:0.268	S6:0.249	S14:0.272	S13:0.289
S10:0.360	S11:0.276	S17:0.286	S1:0.227	S10:0.249	S11:0.247	S7:0.258	S10:0.259
S25							
S25:0.961							
S12:0.306							
S14:0.303							
S11:0.300							
S13:0.295							
S2:0.282							

Table 5.10: AFR-LDA: Rank order of matching scores. The top 6 scene matches and their matching scores are provided for each of the 25 scenes.

AFR-KFA: Rank Order of Matching Scores							
S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
S1:0.993	S2:0.994	S3:0.993	S4:0.992	S5:0.988	S6:0.995	S7:0.996	S8:0.991
S3:0.745	S1:0.729	S1:0.744	S8:0.724	S1:0.660	S10:0.781	S6:0.764	S9:0.773
S2:0.729	S10:0.685	S9:0.722	S14:0.723	S2:0.654	S7:0.765	S10:0.733	S4:0.723
S6:0.681	S9:0.681	S6:0.708	S13:0.686	S6:0.647	S3:0.708	S3:0.640	S14:0.680
S5:0.659	S3:0.659	S10:0.691	S9:0.686	S4:0.634	S9:0.690	S9:0.629	S3:0.668
S9:0.645	S5:0.652	S8:0.667	S11:0.686	S10:0.618	S1:0.681	S4:0.623	S11:0.645
S9	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**
S9:0.992	S10:0.990	S11:0.995	S12:0.996	S13:0.995	S14:0.994	S15:0.994	S16:0.976
S8:0.773	S6:0.781	S13:0.895	S11:0.892	S11:0.895	S11:0.878	S17:0.655	S19:0.709
S3:0.723	S7:0.732	S12:0.891	S14:0.867	S12:0.856	S12:0.868	S19:0.626	S18:0.681
S10:0.693	S9:0.692	S14:0.877	S13:0.855	S14:0.853	S13:0.854	S16:0.607	S17:0.668
S6:0.690	S3:0.691	S4:0.685	S4:0.654	S4:0.685	S4:0.723	S20:0.574	S15:0.624
S4:0.686	S2:0.684	S8:0.645	S8:0.640	S8:0.642	S8:0.681	S18:0.573	S21:0.557
S17	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
S17:0.987	S18:0.985	S19:0.968	S20:0.990	S21:0.985	S22:0.986	S23:0.978	S24:0.978
S18:0.676	S19:0.717	S18:0.724	S21:0.580	S23:0.601	S21:0.537	S21:0.604	S17:0.499
S19:0.663	S17:0.678	S16:0.708	S15:0.575	S18:0.596	S17:0.518	S18:0.591	S21:0.491
S15:0.657	S16:0.673	S17:0.677	S17:0.571	S17:0.595	S5:0.490	S17:0.545	S20:0.476
S16:0.656	S21:0.597	S15:0.648	S18:0.558	S20:0.580	S20:0.481	S25:0.515	S18:0.454
S21:0.590	S23:0.588	S20:0.550	S19:0.533	S6:0.548	S18:0.475	S5:0.504	S16:0.431
S25							
S25:0.983							
S4:0.528							
S5:0.516							
S7:0.514							
S23:0.511							
S6:0.510							

Table 5.11: AFR-KFA: Rank order of matching scores. The top 6 scene matches and their matching scores are provided for each of the 25 scenes.

5.4 Discussion

In this investigation 1 HFR and 3 AFR systems are tested using controlled footage in terms of conveyed information in order to allow a better understanding of how the systems perform. Overall, automated recognition systems are more tolerant to compression than humans. In addition, the performance of HFR and AFR systems with compression is dependent on different face/scene properties. Findings in this investigation have shown compression affects AFR systems performance less with under-exposed ('low lightness') than correctly-exposed ('medium lightness') scenes. This is the opposite for HFR as 'low lightness' scenes were affected the most by

compression and 'medium lightness' scenes the least. Performance of face recognition (AFR and HFR) is scene-dependent and this current investigation proves the importance of including detailed scene properties (derived from scene content characterisation) in face datasets. This will allow exact knowledge on where face recognition systems fail and where they perform the best.

Often the term image usefulness for security applications is used similarly for both humans and automated systems [16], whilst this investigation proves that different scene properties influence recognition systems differently. Perhaps, image usefulness is assessed by defining the system's/process *image acceptance*. *Image acceptance* relates to the acceptability of the scene property/characteristic to complete the recognition task (see Section 3.6).

Additionally, when illumination normalisation techniques are not employed in the evaluation procedure of automated face recognition systems, then the obtained top rank facial images of an 'unknown' face image would entail similar lightness properties to that of the 'unknown' face (i.e. for PCA and KFA methods). This is not the case for the LDA method the as the 'low lightness' category facial scenes were included in the top matches in all the 25 scenes under test. This was unrelated to the face lightness properties of the 'unknown' face.

Furthermore, the performance of the automated systems with 'low lightness' scenes in terms of being affected the least in comparison to the other lightness categories can be explained in many ways. 1) The blacks or low luminance tones are quantised to a greater degree than high luminance tones, indicating that even if the image appears to be predominately black it still contains facial information. 2) The standard compression encoder (MPEGStreamclip) has retained in the compressed scenes the tone characteristics from the reference and the AFR systems (for PCA and KFA methods) have obtained higher scores for scenes within the same lightness category of the 'unknown' face. This perhaps indicates that the AFR systems have performed partner/lightness matching between the dark areas of the 'low lightness' facial images. These dark areas appear to be predominant in the 'low lightness'

scenes as the white areas for the 'high lightness' scenes. If the AFR systems performed pattern/lightness matching, then the 'high lightness' scenes should have obtained similar results to the 'low lightness' category. Only, with the KFA method the 'high lightness' scenes have performed similarly to the 'low lightness' scenes (see Figure 5.7).

CHAPTER 6

Case study 3: The effects of scene content properties, compression and frame rate on the performance of VA systems

Automated tools such as video analytics can be utilised to increase efficiency and usability of the vast amount of CCTV data for the completion of police tasks (see Section 2.1). This has resulted in video analytics becoming a growing area in the security industry. As it has been mentioned in Sections 2.1.2 and 2.1.3, video analytics algorithms often utilise comparable techniques to face recognition algorithms (e.g. detection of faces/human silhouettes from video using segmentation techniques). It is important to understand analytics performance from employing a methodology similar to the AFR systems as described in Chapter 5. Findings are expected to contribute to the development and improvement of both AFR and analytics systems. This current investigation studies the effects of compression and frame rate reduction on the performance of 4 video analytics (VA) systems utilising a low complexity scenario. Additionally, the most influential scene properties affecting the performance of these systems are identified.

6.1 Introduction

The image library for intelligent detection systems (iLIDS) provides datasets with various scenarios of video surveillance. This is a UK government initiative for the development and selection of VA systems. Each scenario is made up of 3 datasets: 2 publicly available (training and test datasets) and 1 privately held evaluation dataset. The private one is used in order to benchmark the performance of VA systems and provide the developers with a UK Government classification standard [50]. Part of the publicly available Sterile Zone (SZ) dataset of iLIDS scenarios is investigated in this paper. The SZ is a low complexity scenario, consisting of a fence (not to be trespassed) and an area with grass (see Figure 6.1). The VA system needs to alarm when there is an intruder entering the scene (an attack). The iLIDS datasets can be obtained from the Home Office Centre for Applied Science and Technology (CAST), to assist those wishing to investigate solutions in relation to the VA systems [49].

Figure 6.1: Example camera views from the iLIDS dataset. The subjects in the footage wear only 2 types of clothing: white or green.

The aim of this investigation is to identify the effects of compression and reduction of frame rate on the performance of 4 VA systems with the SZ scenario. Furthermore, to identify the most influential scene properties affecting the performance of each VA system under investigation.

The 4 VA systems under investigation have obtained UK Government approval by being tested with analogue DigiBeta videocassettes at D1 PAL resolution (720×576). This information has been provided by CAST, they have highlighted that no further information can be provided in relation to the systems in order to protect the manufacturers' product and privacy. This chapter includes investigative work relating to the testing of 4 VA systems with D1 PAL resolution of uncompressed and compressed (6 levels of compression using H.264/MPEG-4 AVC MPEG Streamclip encoder at 25 and 5 frames per second) footage, consisting of quantified scene properties. The scene properties were extracted from the characterisation of the content of 110 attacks (scenes). The characterisation included both objective and subjective techniques relating to scene contrast (contrast between main subject and background), camera to subject distance, subject description (e.g. 1 person, 2 people), subject approach (e.g. run, walk), and subject orientation (e.g. perpendicular, diagonal). After the characterisation, the scenes were grouped based on common properties. Additional footage, including only distractions (i.e. no attacks to be detected) is also investigated. Distractions are elements in the scene such as abrupt illumination changes and birds that could be falsely recognised by the systems as intruders.

Section 6.2 presents the experimental methodology. Data analysis and discussion of the results are described in Section 6.3. Lastly, in Section 6.4, conclusions are drawn.

6.2 Methodology

The methodology included 3 main steps: a) preparation of the test footage (uncompressed and distorted), b) scene content characterisation to define image properties, and c) testing of the VA systems.

6.2.1 Preparation of the test footage

The SZ dataset is segmented into shorter video clips. Table 6.1, provides a general description of the 17 clips under investigation. These clips include 110 attacks and have 11 hours duration of footage. This part of the dataset was selected based on the availability of the original tape recordings of the scenario. The uncompressed footage was originally recorded using analogue DigiBeta videocassettes at D1 PAL resolution (720×576), 50ifps (interlaced frames per second) and a bitrate of around 90 megabits per second (Mbits/s). DigiBeta uses a lossless compression at 10-bit, compressing YUV channels with a chroma subsampling corresponding to 4:2:2. The iLIDS team provides the publically available datasets with 10% compression and only the tapes could have been used to obtain the uncompressed reference.

The original videocassettes were digitised using the Apple^{TM} Final Cut Pro^{TM} (FCP) uncompressed format. The FCP uncompressed format uses similar specifications to DigiBeta: 8bit YUV 4:2:2 and 96 Mbits/s bitrate. Furthermore, all clips were de-interlaced in FCP by removing one of the odd fields and interpolating the even number of fields in order to avoid any problem with the interlaced effect when transmitting the video clips to the VA systems. This should not affect the results, as the VA systems would grab the fields to further analyse (based on how analogue signals behave) rather than the progressive frames. Thus the reference original in this investigation is in FCP uncompressed format at 96 Mbits/s, and at 25 progressive frames per second. The MPEG Streamclip implementation encoder was employed to compress the clips at selected target bitrates and frame rates using the video coding standard H.264/MPEG-4 AVC, which is widely used in surveil-

Clip name	Attacks	Duration	Time	Weather	Distractions
1) sztea101a	10	00:37	Dawn	None	Camera switch from monochrome to colour and opposite
2)sztea101b	15	00:49	Dusk	None	Camera switch from monochrome to colour, bats
3)sztea102a	13	00:37	Dawn	None	Camera switch from monochrome to colour
4)sztea102b	14	00:46	Day	Overcast	Vehicle
5)sztea103a	17	00:47	Day	Clouds	None
6)sztea104a	31	01:32	Night	None	Bats
7)sztea105a	10	00:35	Day	Overcast, Snow	None
8)szten101a	none	00:15	Day	Overcast	Bag, squirrel, small illumination variations
9)szten101b	none	00:30	Day	None	Rabbits, shadow through fence, illumination variations
10)szten101c	none	00:30	Dusk	None	Camera switch from colour to monochrome, birds, rabbits
11)szten101d	none	00:30	Dawn	None	Birds, rabbits, illumination variations
12)szten102a	none	00:45	Day	Some	Birds, illumination variations, shadow through fence
13)szten102b	none	00:30	Day	Overcast, Rain	Birds, small illumination variations
14)szten102c	none	00:30	Day	Overcast, Snow	None
15)szten102d	none	00:15	Dusk	Overcast	Camera switch from colour to monochrome, foxes, rabbits
16)szten103a	none	00:40	Night	None	Small changes of camera positioning because of wind
17)szten103b	none	00:30	Day	Overcast	Small changes of camera positioning because of wind

Table 6.1: Part of the SZ scenario dataset under test. The table provides information in relation to the general description of each clip. The first 7 clips contain attacks and the last 10 clips contain only distractions. Most of the information has been obtained from the ground truth dataset.

lance applications (see Section 3.1). The MPEG Streamclip encoder was selected with only bitrate control (i.e. no GOP size or B frames were selected), because it complies with the common functioning of security recording systems which are based only on bitrate control (see Section 4.2.3). The compression bitrates used were approximately the following in kilobits per second (kbps) for each type of the chosen frame rate:

- 25fps: 200, 400, 800, 1200, 1800, and 2000;
- 5fps: 40, 80, 160, 240, 320, and 400.

The degraded footage produced at 5fps repeats each of the extracted 5 frames, 5 times in each second. The range of the bitrates at 5fps were chosen to be equivalent to the bitrates at 25fps taking into consideration the reduction of frame rate. For example, 2000kbps at 25fps would be 400kbps when reducing the frame rate to 5fps (i.e. $\frac{2000\times5}{25} = 400$). The test footage for the VA systems consists of the reference and its 12 degraded versions. The range of the degraded versions was chosen to cover a variety of compressed qualities (high and low).

6.2.2 Scene content characterisation

The characterisation of the scene content of each attack is carried out to enable a better understanding of the properties that might affect the performance, in terms of correct detection, of analytics systems. The influential properties could be related to image quality attributes (e.g. contrast or sharpness), and/or the properties of the subject to be detected (e.g. orientation). Each of the 110 attacks (or footage of attacks) was classified into content properties.

Table 6.2 includes the names and total number of each property in each grouped category. The properties and groups relating to the description of the subject (groups: approach, description, distance, and orientation) in the attacks were extracted by visual examination (apart from the distance group) and were already available within the ground truth data of the SZ dataset. The approach group properties describe the way the subject approaches the fence and consists of 9 lev-

els (e.g. walk, run - refer to Table 6.2). The description group properties consist of 2 levels and explain if the subject includes 1 person or 2 people next to each other (i.e. 2 people in the scene indicates a bigger subject area to be detected). The distance group properties consist of 3 levels and describe the distance of the subject to the camera; far -30 meters away from the camera, middle -15 meters away from the camera, and close -10 meters away from the camera. Figure 2.9 provides an example of the distance group properties. Orientation group properties consist of 2 levels and indicate if the attack happened perpendicular or diagonal to the fence. If the attack happens diagonally then the subject is in the scene for a longer time than with a perpendicular attack.

Properties describing the image quality of the attack belong to the contrast group and their values were obtained by simply using a contrast ratio (CR - see Eq. 3.11) of dark to light area between foreground (attacker) and background (grass area). Refer to Section 3.5.1 for more information on the contrast ratio characterisation technique. The lightness values in the CR calculations were derived by measuring lightness in specific areas in the scene using the CIELab L^* colour space (see Section 3.3.4). For each attack scene, 2 lightness measures were derived: 1) on the surrounding grass area of the subject/s (the average of 4 areas around the attacker: above, below, left and right), and 2) on the clothing of the subject/s (the average of 4 areas on the attacker: upper body, lower body, left and right legs). The subjects, in the footage wear only 2 types of clothing: white or green. The head of the subject/s was excluded from the measurements in order to avoid complications with the measured lightness. Furthermore, these measurements were applied on 3 different positions of the attacker in the scene (beginning, middle and near to the fence). The mean value of the 3 positions in the scene was selected to be used in the CR formula. In Table 6.2, next to the properties under the contrast group, information on the range of the obtained contrast values is provided along with the total number of scenes for each property. Furthermore, the green clothing incorporates camouflage properties and the white clothing perhaps can be more distinctive from the background (grass area) in comparison to the green clothing. It is unknown what

Approach	**Description**	**Distance**	**Orientations**
1. Walk: 28	1. One person: 98	1. Far: 36	1. Perpendicular: 98
2. Run: 20	2. Two people: 12	2. Middle: 37	2. Diagonal: 12
3. Creep walk: 15		3. Close: 37	
4. Crawl: 11			
5. Crouch Run: 11			
6. Crouch Walk: 9			
7. Body Drag: 7			
8. Log Roll: 3			
9. Walk with ladder: 6			
Contrast			
1. Low(0.00-0.33): 23			
2. Medium(0.33-0.66): 78			
3. High(0.66-0.99): 9			

Table 6.2: Properties and grouping of scenes. Each column provides the group identification name (e.g. contrast), its properties (e.g. low) and the total number of scenes in each property.

would happen in a real case where the intruder might wear another colour.

6.2.3 Testing of the VA systems

The 4 VA systems under investigation were isolated units (not incorporated within a recorder) and were designed to take composite signals as input. The VA systems have been optimised for the testing with the iLIDS SZ scenario. The 4 systems have received UK Government approval and could be further classified as operationally successful systems (see Section 6.1). The systems have been labeled as A, B, C and D.

For measuring the performance of the analytics systems, a method was required to simultaneously play the video clips and record the alarm attacks raised. Important criteria were to keep the video quality as high as possible and the ability to accurately determine the time-code from the video file, so that alarm times could be recorded precisely. The VLC application from VideoLan [288] was chosen to act as the player running on an Intel i7 PC with Windows 7.

An ATI Radeon X1300 graphics card with PAL composite output was used to feed the analytics systems via a Kramer 105VB distribution amplifier (see Figure 6.2). A broadcast standard graphics card was considered, but the effort to integrate this with the system was beyond the scope of the project. The analytics systems signal the detection of an alarm attack by shorting out a normally open contact on 1, or more of their output connectors. To interface these to the PC, an Amplicon PCI236 Digital I/O card was used via an EX230 Isolation Panel. A bespoke software application written in C# was used to integrate VLC with the Amplicon card. The Net API called nVLC [289] was used to interface to the VLC libraries directly and derive a accurate time-code from the playing video.

The developed software allows for multiple video clips to be queued for play-out, with each clip being able to play multiple times. With a video clip playing, alarm attacks were captured via the Amplicon card. Each alarm was saved along with the

corresponding clip time, clip name, system name and repeat number to a simple text file. The ground truth data for each clip was then compared with this file.

In theory there should be no latency and if any it should be minimal (1 to 2 seconds). The software that has been created reads the time-code directly from the VLC player and uses this information to time-stamp the relayed triggers. The Amplicon card communicates the triggers to the software and the software records the end results in an Excel spreadsheet. To test the system between time-stamp and recording of results a manual trigger (switch) was utilised and no latency issues were detected. Also, no latency issues were detected from visual observations between playing the video(s) and recording of results.

The rules determining whether an alarmed attack was true or false were defined as follows: if an alarm falls within the ground truth alarm period, then a true match is recorded; if there are further alarms within the same period they are ignored; if an alarm occurs outside of the ground truth period, then that is noted as a false alarm. The obtained results have scores of 1 for the correctly detected attacks and 0 for the missed attacks. To estimate the consistency of recording the results, each clip was repeated 10 times with black video of 30 seconds played between each clip to reset the algorithm settings. Most of the manufacturers of the systems confirmed that it takes about 10 seconds for their algorithms to be adjusted/tuned to the scene content (e.g. weather conditions such as rain or snow).

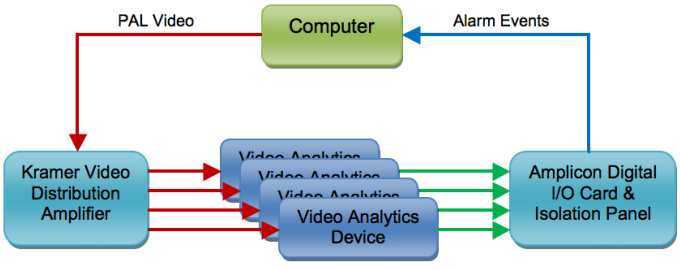


Figure 6.2: Video distribution and recording of results.

There were some small variations in the results between the repeats due to the noise added to the video signal (i.e. as part of the output of footage to the detection

systems), and/or the actual intrinsic parameters of the video analytics systems (i.e. how it is tuned/designed) and/or the properties of the events (i.e. it was observed that variation was triggered by certain events). In fact many statistical algorithms that could be incorporated within the VA systems under test could utilise randomisation techniques to initially decide where spatially to put clusters, nodes or other-kinds-of-data. A typical example is the K-means technique in which the selection of initial cluster-centres is often randomised and does not provide an optimum result [290]. This phenomenon was investigated further by repeating the whole process (10 repeats on 3 clips with attacks) another 5 times. The derived proportional values (i.e. average of 10 repeats) among the further 5 repeats were consistent and similar.

6.3 Results

The analysis of the results has been divided into 3 parts. The first part identifies the overall detection performance for each individual system with respect to compression (Section 6.3.1); the second part identifies the most influential attack/scene properties for each individual system with respect to compression (Section 6.3.2); and the third part provides an analysis on false alarms (Section 6.3.3).

6.3.1 Overall detection performance analysis with respect to compression

The overall detection performance investigates the output of the VA systems for all the attacks with respect to compression (at 25fps and 5fps). As mentioned in Section 6.2.3, all the VA systems under investigation have produced some variation in the results from the 10 repeats of each clip/attack. In Figures 6.3 and 6.4, each row corresponds to one of the VA systems A, B, C and D. From the left to right columns, the graphs depict:

a) the points (connected with a line) representing the proportion of the always

correctly identified scenes from all the 10 repeated trials (the *Yes* scenes) plotted against the different levels of compression (i.e in kbps) and the uncompressed reference,
b) the points (connected with a line) representing the proportion of the always missed scenes from all the 10 repeated trials (the *No* scenes) plotted against the different levels of compression (i.e in kbps) and the uncompressed reference, and
c) the points (connected with a line) representing the proportion of the uncertain scenes; those that have produced varied detection from the 10 repeated trials (the *Uncertain* scenes) plotted against the different levels of compression (i.e in kbps) and the uncompressed reference.

The sum of the corresponding proportions in the 3 graphs (*Yes*, *No* and *Uncertain* scenes) would be equal to 100%.

The results in Figures 6.3 and 6.4 have shown that every system performed differently for each compression/frame rate level (see the *Yes*, *No* and *Uncertain* scene graphs), but overall compression has not adversely affected the performance of the systems. Some systems have performed better (A and D) than others (B and C). For example, in Figures 6.3 and 6.4 the total number of attacks always detected (*Yes* scene graphs) is higher for both 25fps and 5fps for the better systems than the rest. Some further observations can be made from the *Yes* scenes graphs: a) System A performance has dropped with reduced frame rate and high compression levels (200kbps at 25fps and 40kbps at 5fps), b) System B performance has dropped with the reference footage and a slight increase can be seen at 2000kbps with 25fps. Also, performance has dropped with reduced frame rate and with higher compression at 5fps. c) System C performance seems to be constant throughout the different levels of compression/frame rates and an increase of performance can be seen at higher compression levels (200kbps at 25fps) and with reduced frame rate. d) System D performance has dropped with reduced frame rate and high compression levels at 5fps (40kbps at 5fps).

In the *No* scenes graphs, the performance at 25fps and 5fps was similar for systems

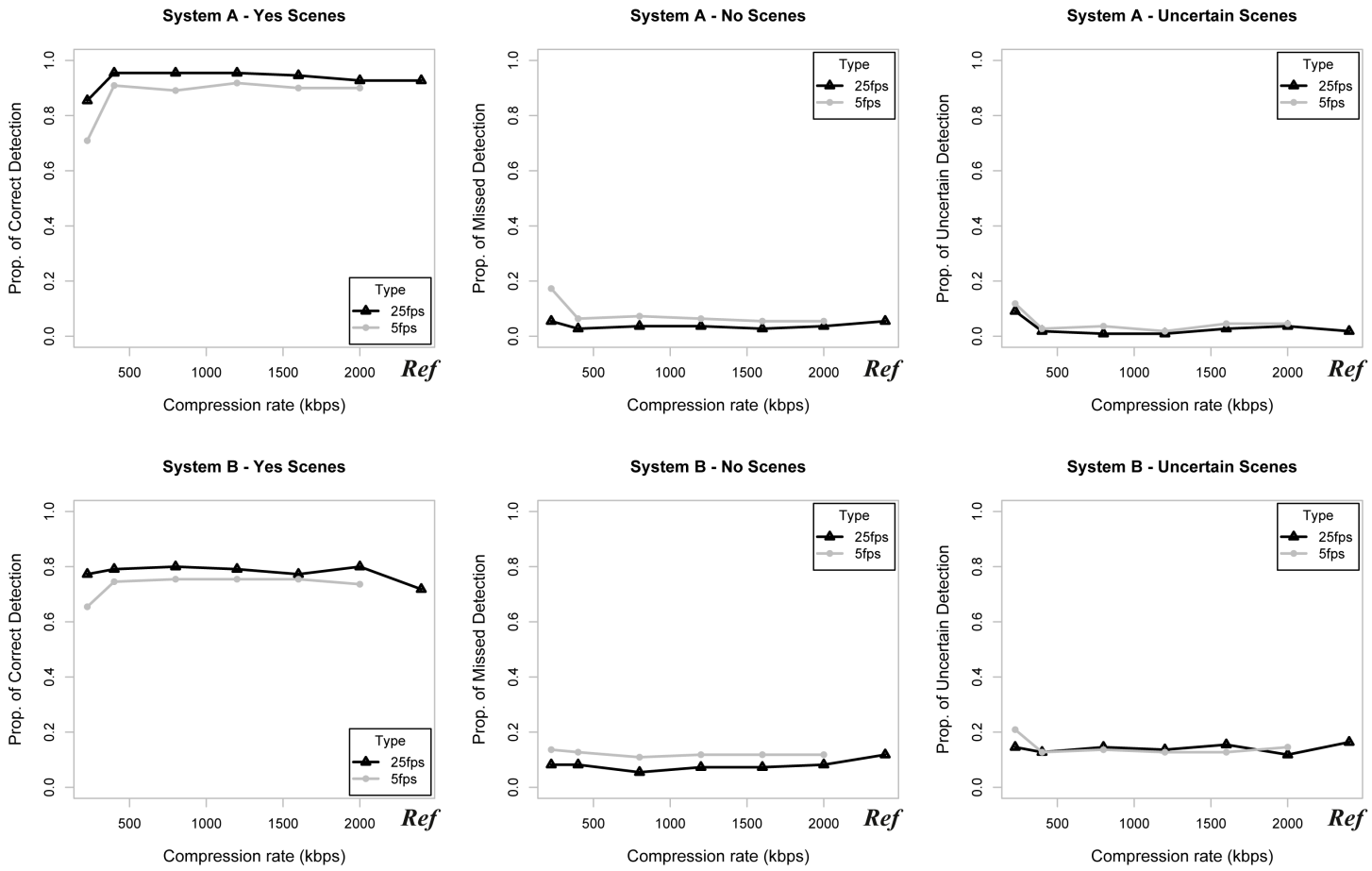


Figure 6.3: Overall detection performance with respect to compression for systems A and B. Black triangles and black lines represent derived results from 25fps, and grey dots and grey lines represent derived results from 5fps. The points of the proportion of identified attacks (connected with a line) for each type of scene (*Yes*, *No*, *Uncertain*) are plotted against the kbps and the reference.

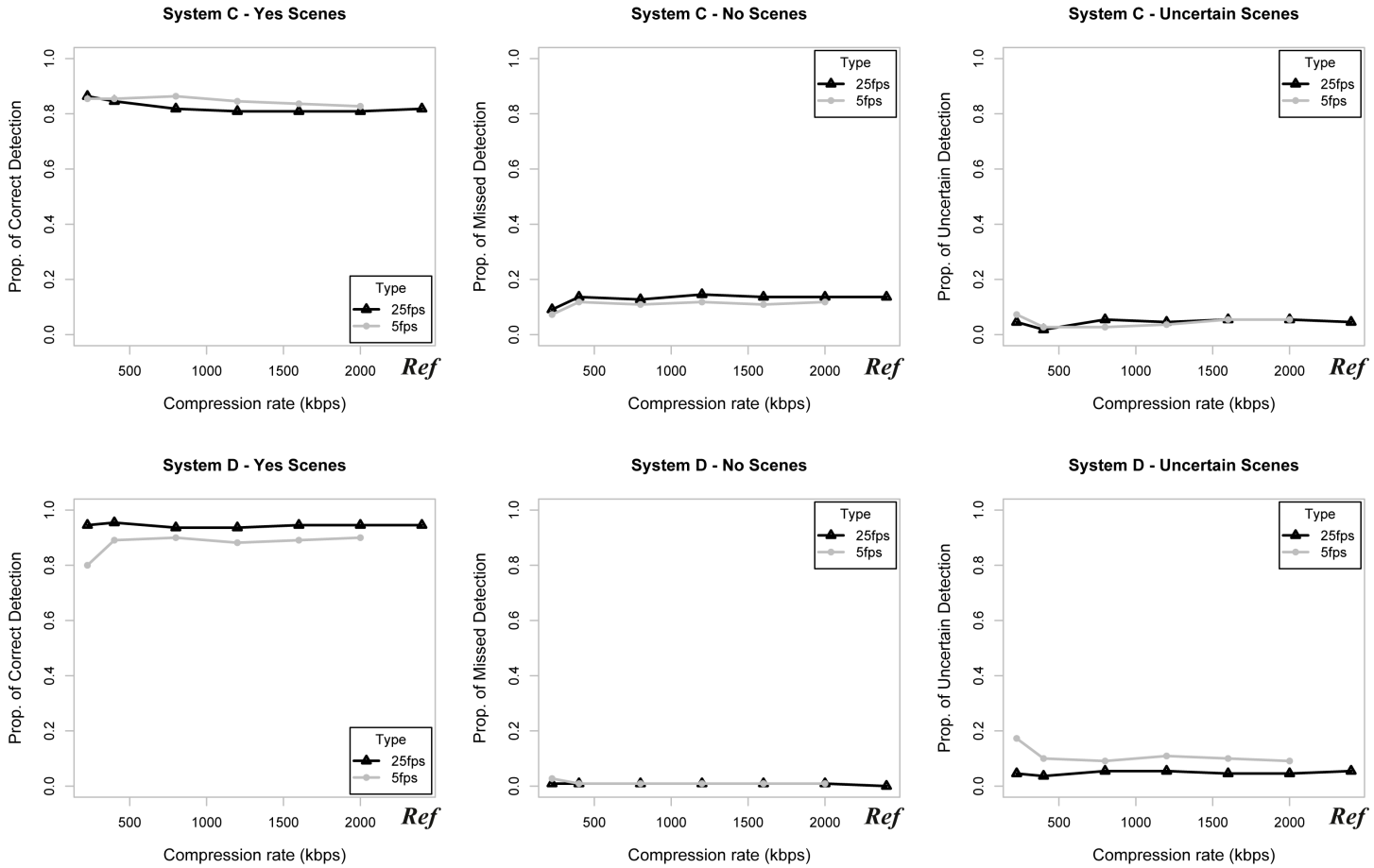


Figure 6.4: Overall detection performance with respect to compression for systems C and D (as graphs in Figure 6.3).

C and D (in Figure 6.4). For systems A and B (in Figures 6.3), more missed attacks were observed at 5fps. Additionally, in the *Uncertain* scenes graphs, the performance at 25fps and 5fps has been the same for systems A, B and C. A drop in performance, in terms of proportion of attacks causing uncertainty, can be seen for system D at 5fps.

All graphs in Section 6.3.1 and Section 6.3.2 consist of raw points (black triangles and grey dots) and lines (black and grey, most of them are logistic regression models), which correspond respectively to results obtained from 25fps and 5fps.

The aforementioned analysis has basically categorised the results into 3 groups which are: success (always correct detection-score of 1), failure (always missed detection-score of 0) and uncertainty (both correct and missed detection-scores ranged between 0.1 to 0.9). The derived results are not strictly binary but rather proportional (i.e. by utilising the 10 repeats of each clip/attack) with a binary nature of success or failure. 2 further approaches (logistic regression and linear regression) of analysing such data have been identified and implemented that take into consideration compression amounts/levels and are described by the following paragraphs. Further, both these approaches provide visual understanding (i.e. plots) of the obtained data. An additional factor analysis was considered but proven to be inappropriate for the type of data derived from this investigation. For example, a factor analysis does not take into consideration the different levels of compression for each scene property (see Section 6.3.2) and the method produced complicated/unrealistic models (i.e. the scene properties under investigation are too many-19). The following paragraphs provide information on the regression models implemented for both the overall (i.e. all scenes/attacks are included) and detailed (i.e. analysis is based on individual scene properties - see Section 6.3.2) analysis of detection performance.

1. Logistic Regression. In order to take into consideration the number of successes (and failures) in an n repeated number of trials (i.e. in this case 10) for each attack, the recorded results were modelled using logistic regression with

the generalised linear model (gml) function in R software for statistics [282]. In this way a weighted regression is carried out (all the *Yes*, *No*, and *Uncertain* scenes are taken into consideration), using the number of trials as weights and the logit-link function to ensure linearity [291–293]. The logit-link function is a transformation that uses the natural log of odds; the ratio of 2 probabilities (see Eq.6.3). This is how it is derived, Eq. 6.1 provides a logistic model of proportion P as a function of x (which normally produces an S-shaped curve). This logistic model is linearised by substituting the proportion P with the odds p/q; where p are the successes and where q the failures (see Eq. 6.2). Finally, the linear predictor is obtained for the odds by taking the natural log (see Eq. 6.3). Where p/q is the response variable, x the explanatory variable, a the intercept, and b the slope. Even though the result is the linearisation of a logistic function, the fit of a linear model needs to be avoided as neither the normality or the homoscedasticity assumptions are met [292]. The parameters and error estimates for a logistic regression analysis are derived via maximum likelihood and for a linear regression analysis via least square. In this Section 6.3.1 and Section 6.3.2 the results are modelled using logistic regression.

$$P = \frac{e^{(a+bx)}}{1 + e^{(a+bx)}} \tag{6.1}$$

$$\frac{p}{q} = \frac{e^{(a+bx)}}{1 + e^{(a+bx)}} \left[1 - \frac{e^{(a+bx)}}{1 + e^{(a+bx)}}\right]^{-1} = e^{(a+bx)} \tag{6.2}$$

$$ln\left(\frac{p}{q}\right) = a + bx \tag{6.3}$$

2. Linear Regression. This is the simplest form of regression and a model is fitted in order to understand the relationship between 2 continuous variables (i.e. consisting of real numbers) [283]. Eq.6.4 provides a simple linear model that has been employed in this analysis; where y is the response variable, x the explanatory variable, a the intercept, and b the slope. This linear model has been fitted only to the proportion of the always correctly detected scenes (the *Yes* scenes) and the results are presented in Appendix C. This method

has only been accomplished for comparison reasons. The derived data from this investigation are better fitted under the logistic regression analysis (see above paragraph on logistic regression approach).

$$y = a + bx \tag{6.4}$$

All the fitted regression models, in Section 6.3.1, Section 6.3.2 and in Appendix C were carried out in R software for statistics. Figure 6.5 presents the derived logistic regression models obtained for each VA system (A, B, C and D) from modelling all the raw data with respect to the different levels of compression (i.e. in ln kbps). Displaying all the raw data (as proportion of correct detection) in the graphs causes confusion as they are spread disproportionally around the graphs; instead the proportion of the always correctly identified scenes from all the 10 repeated trials (the *Yes* scenes) are plotted. The majority of the obtained results do belong in the *Yes* scenes category and as a result the fitted logistic regression models will be close to the proportion of the *Yes* scenes points but not exact models of them. In the fitting of the logistic regression models, all the scenes (*Yes*, *No* and *Uncertain*) are taken into consideration. Additionally, as the derived data from this investigation seem to exhibit a complicated behaviour (e.g. *Yes*, *No* and *Uncertain* scenes) then all the analysis that has been carried out has the purpose of understanding trends rather than fitting the perfect models.

Table 6.3 includes details of the fitted logistic regression models in Figure 6.5. The first column provides the system name and the type of raw data (25fps or 5fps). The second and fourth columns provide information on the derived coefficients of each model (intercept and slope). Their next columns provide the calculated standard error on the coefficients (std). Where p is the statistical value identifying any significant trends. For example, the results in Table 6.3 indicate a significant correlation between proportional detection of attacks and compression levels for only systems A and D, and at only 5fps. We conclude that the proportion of correct attack detection for systems A and D at 5fps increases significantly with increas-

ing kbps (less compression). For the rest of the compression levels and systems, compression has not affected the overall performance of the systems (see Figure 6.5 and Table 6.3). This is overall a positive result, since it indicates that footage can be significantly compressed (for storage or transmission purposes) with very little reduction in correct attack detection.

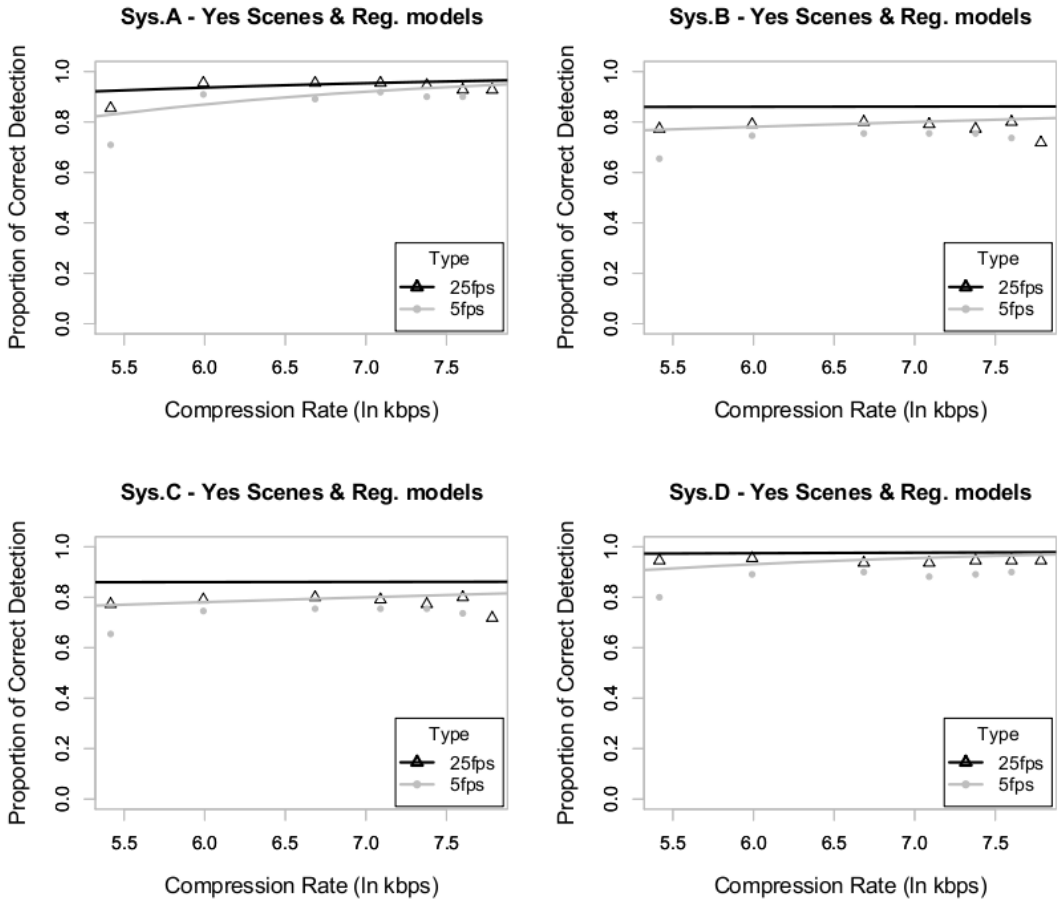


Figure 6.5: Overall performance with respect to compression (in ln kbps) for systems A, B, C and D. Black triangles and black lines represent derived results from 25fps, and grey dots and grey lines represent derived results from 5fps. The lines are the obtained logistic regression models from all the scenes and the points represent the always correctly identified scenes (the *Yes* scenes), both plotted against the natural logarithm of compression rate in kbps.

Appendix C provides the results from the linear analysis for the overall performance of the VA systems under test. The findings from the linear regression (Figure C.1 and Table C.1) are similar to those from logistic regression in terms of compression not affecting the performance of the systems; except for system D at 5 fps where there is a significant correlation between proportion of the *Yes* scenes and

compression levels (performance decreases as compression amount increases).

System	α	**std**	β	**std**	p
$Sys.A_{25fps}$	0.665	1.365	0.337	0.208	0.106
$Sys.A_{5fps}$	-1.376	1.026	0.546	0.158	0.000***
$Sys.B_{25fps}$	1.782	0.893	0.006	0.133	0.966
$Sys.B_{5fps}$	0.579	0.761	0.115	0.114	0.311
$Sys.C_{25fps}$	2.863	0.961	-0.166	0.141	0.242
$Sys.C_{5fps}$	2.829	1.005	-0.139	0.148	0.347
$Sys.D_{25fps}$	3.084	1.804	0.095	0.269	0.726
$Sys.D_{5fps}$	-0.172	1.079	0.463	0.166	0.005**

Table 6.3: Information of the fitted logistic regression models in Figure 6.5 for the overall performance. The first column provides the system name and the type of the raw data (25fps or 5fps). The second and fourth columns provide information on the derived coefficients of each model (α-intercept and β slope). Next columns provide the calculated standard error on the coefficients (std). Where p is the statistical value identifying any significant trends (signif. codes: 0‘***’, 0.001 ‘**’, 0.01‘*’).

6.3.2 Detailed performance analysis with respect to compression

This section includes diagnostics in terms of providing a detailed analysis on the performance of each system for each scene property under investigation. An analysis based on the scene content properties enables understanding on where systems need improvement. Figures 6.6 and 6.7 correspond to system A, Figures 6.8 and 6.9 correspond to system B, Figures 6.10 and 6.11 correspond to system C, Figures 6.12 and 6.13 correspond to system D and each pair of figures (e.g. Figures 6.6 and 6.7) includes graphs for each of the 19 scene properties under investigation (see Table 6.2). In these graphs, the lines are the obtained logistic regression models for 25fps (black line) and 5fps (grey line) (similarly to the graphs in Figure 6.5 in Section 6.3.1). When the lines in the graphs overlap, the grey line is on top and the black line becomes invisible. The vertical axes represent proportion detection (ranging between 0 and 1) and the horizontal axis the natural logarithm of compression rate (in ln kbps). Figures 6.6 to 6.13 provide a visual understanding on how each system has performed for each scene property; detailed information of the fitted logistic regression models is provided in Appendix B (i.e. coefficients and

their calculated standard errors, and the statistical p value identifying any significant trends). Furthermore, good performance can be visualised when most of the logistic regression models are near to maximum detection (vertical axes value of 1). Reduction of performance can be visualised when the logistic regression lines incline to minimum detection (vertical axes value of 0).

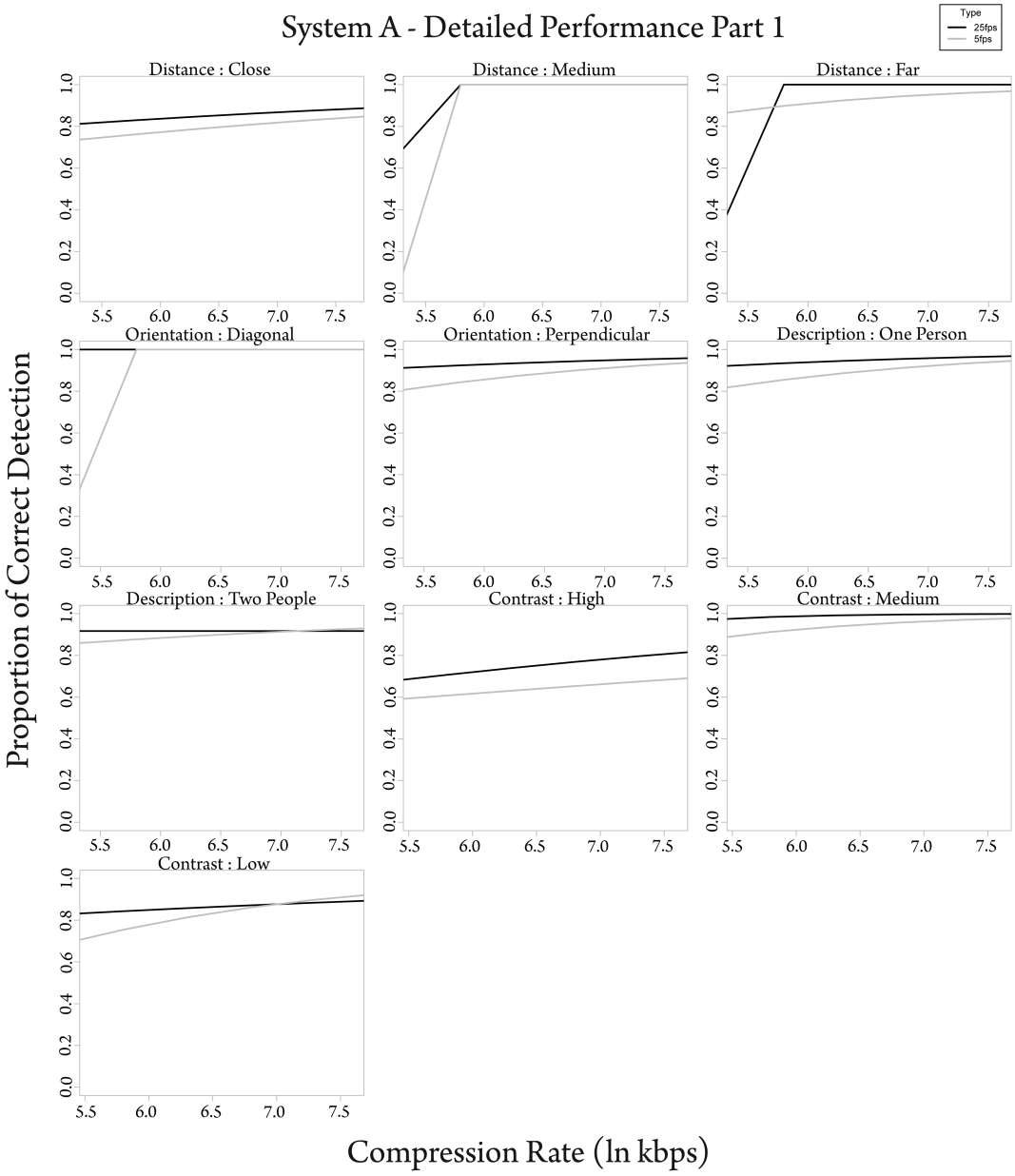


Figure 6.6: Detailed performance with respect to compression (in ln kbps) for system A Part 1. Black lines represent derived results from 25fps and grey lines represent derived results from 5fps. The lines are the obtained logistic regression models for each individual scene property plotted against the natural logarithm of compression rate. The vertical axes represent proportion of detection, ranging between 0 and 1.

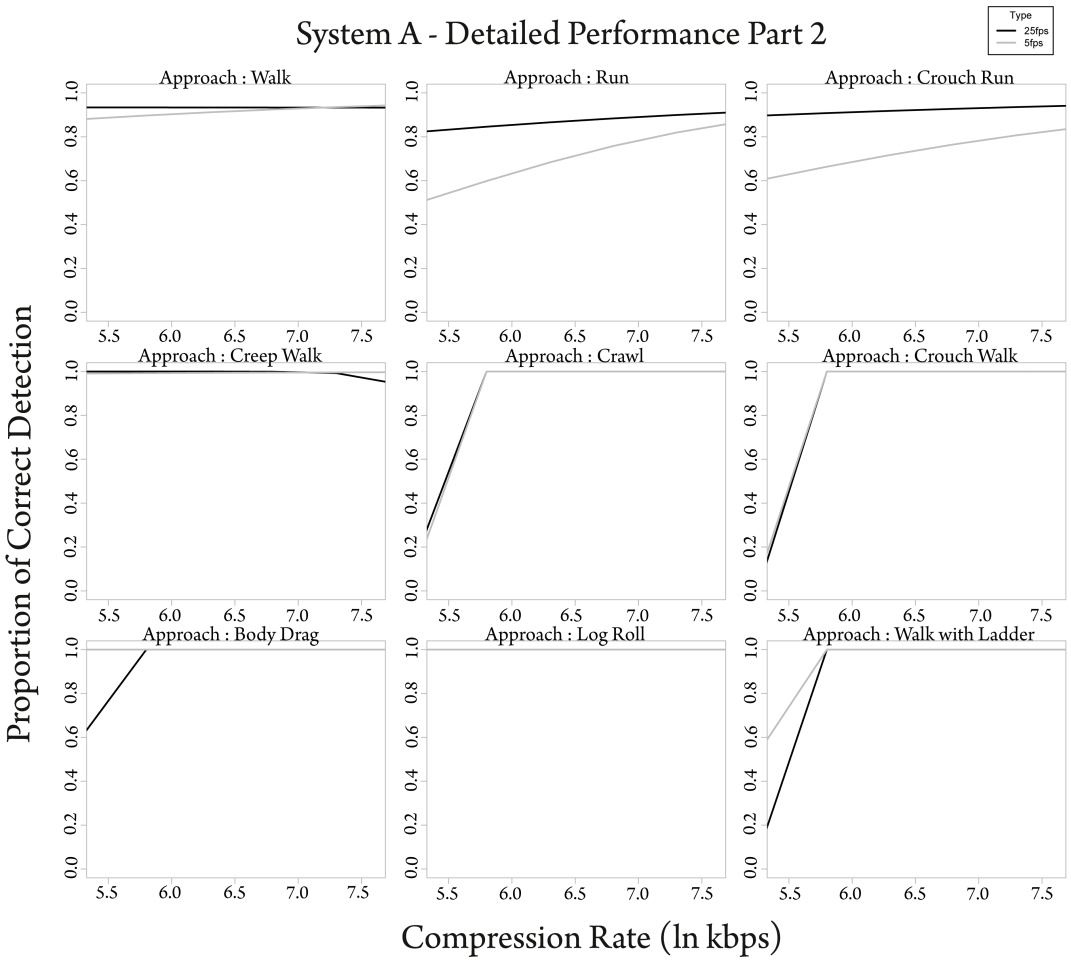


Figure 6.7: Detailed performance with respect to compression (in ln kbps) for system A Part 2 (as graphs in Figure 6.6).

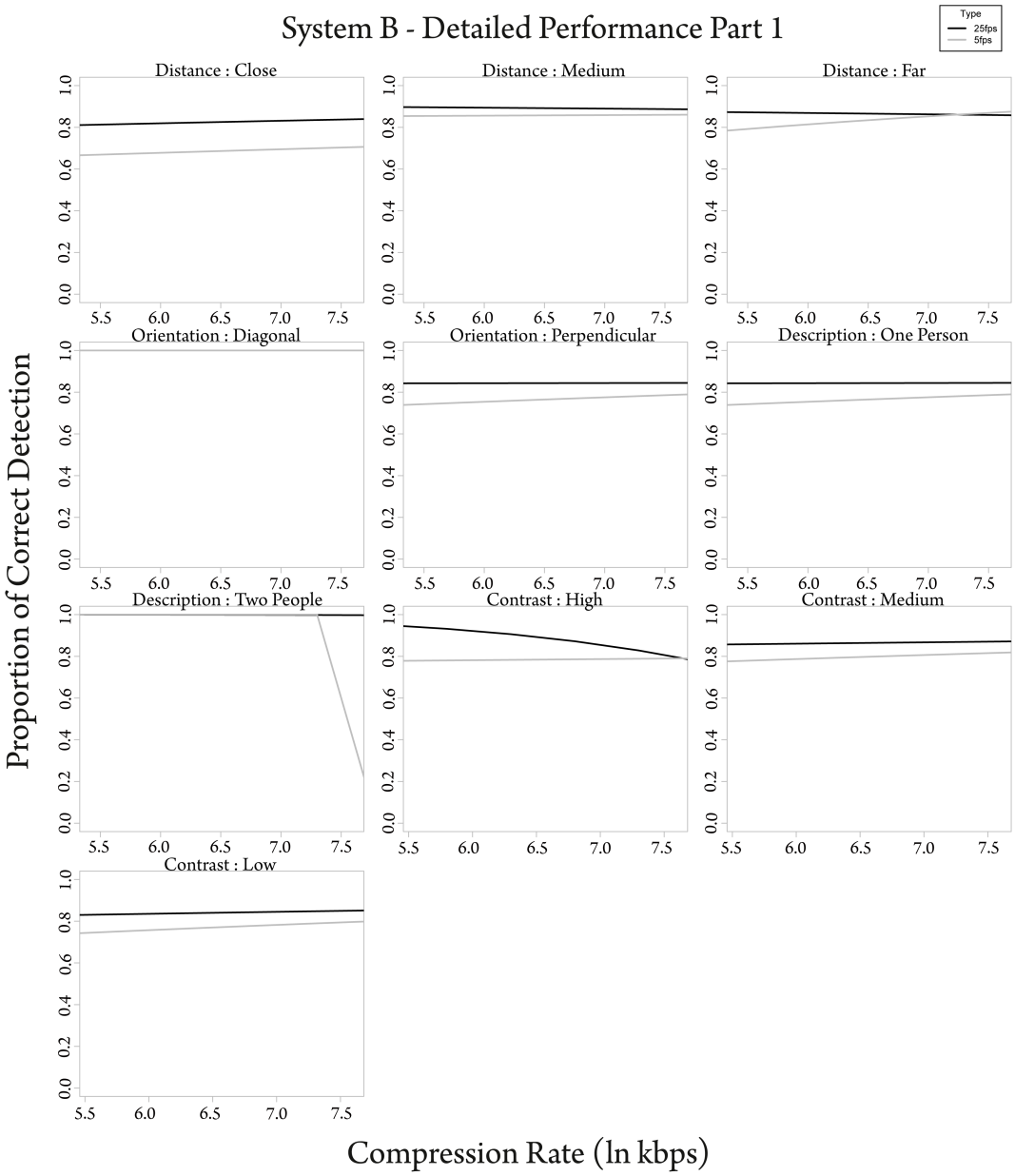


Figure 6.8: Detailed performance with respect to compression (in ln kbps) for system B Part 1 (as graphs in Figure 6.6).

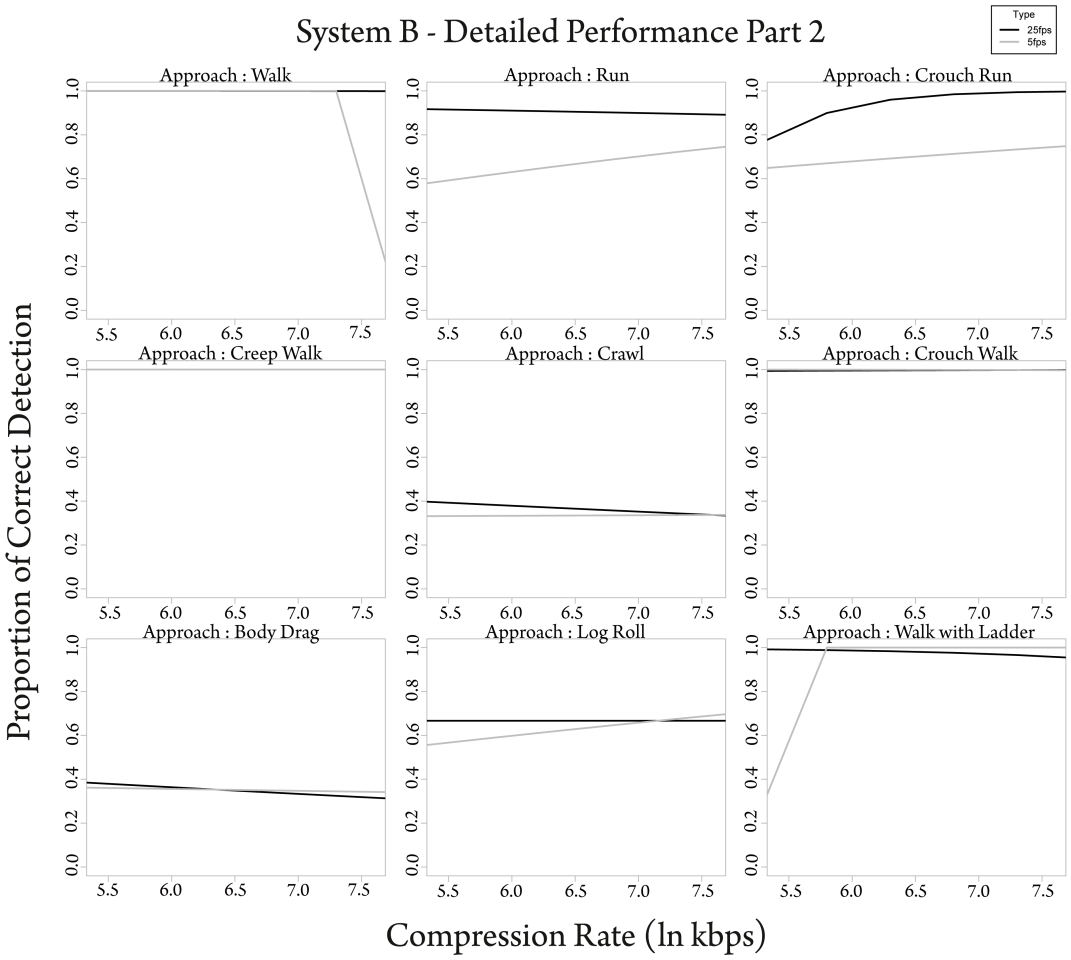


Figure 6.9: Detailed performance with respect to compression (in ln kbps) for system B Part 2 (as graphs in Figure 6.6).

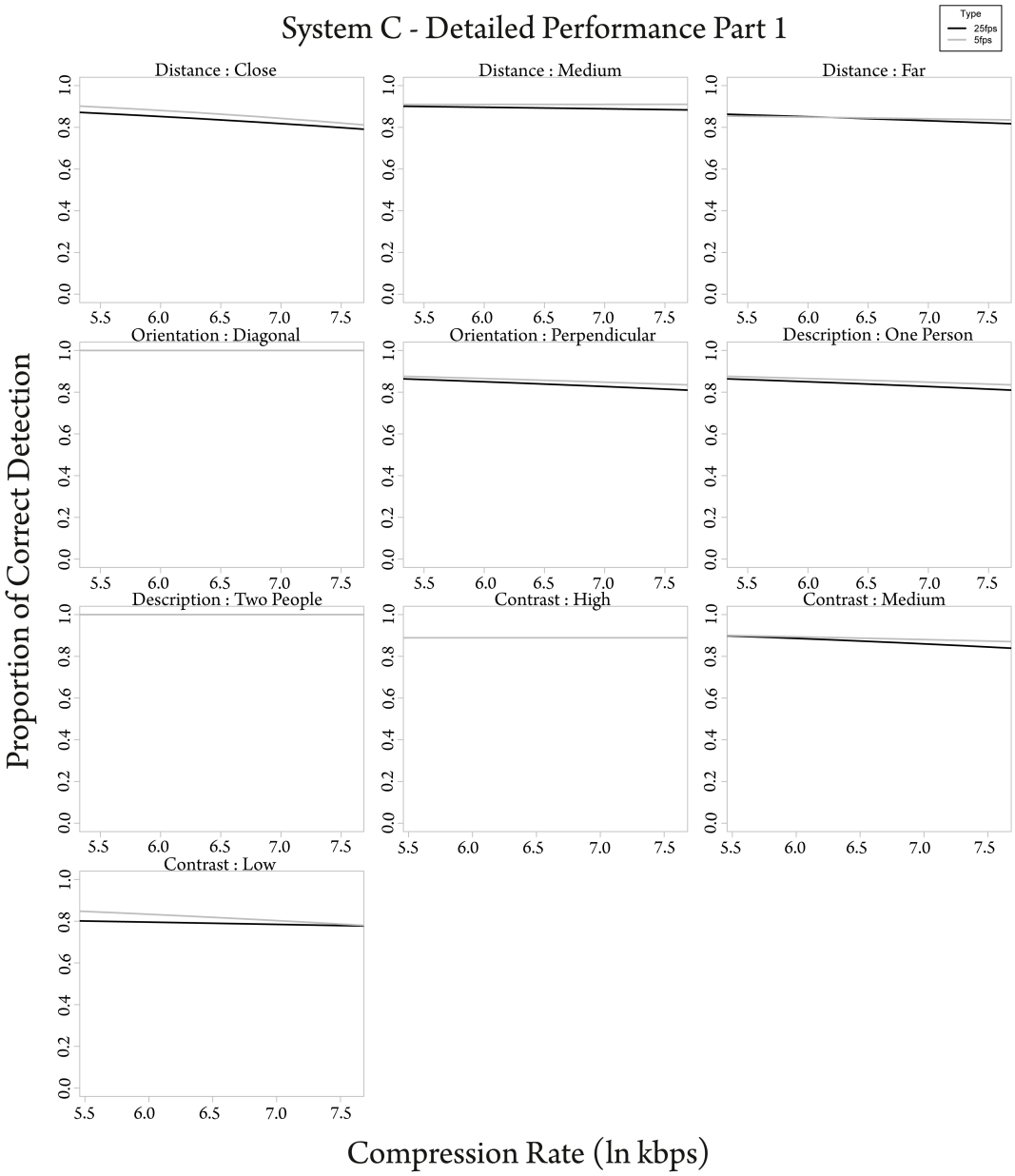


Figure 6.10: Detailed performance with respect to compression (in ln kbps) for system C Part 1 (as graphs in Figure 6.6).

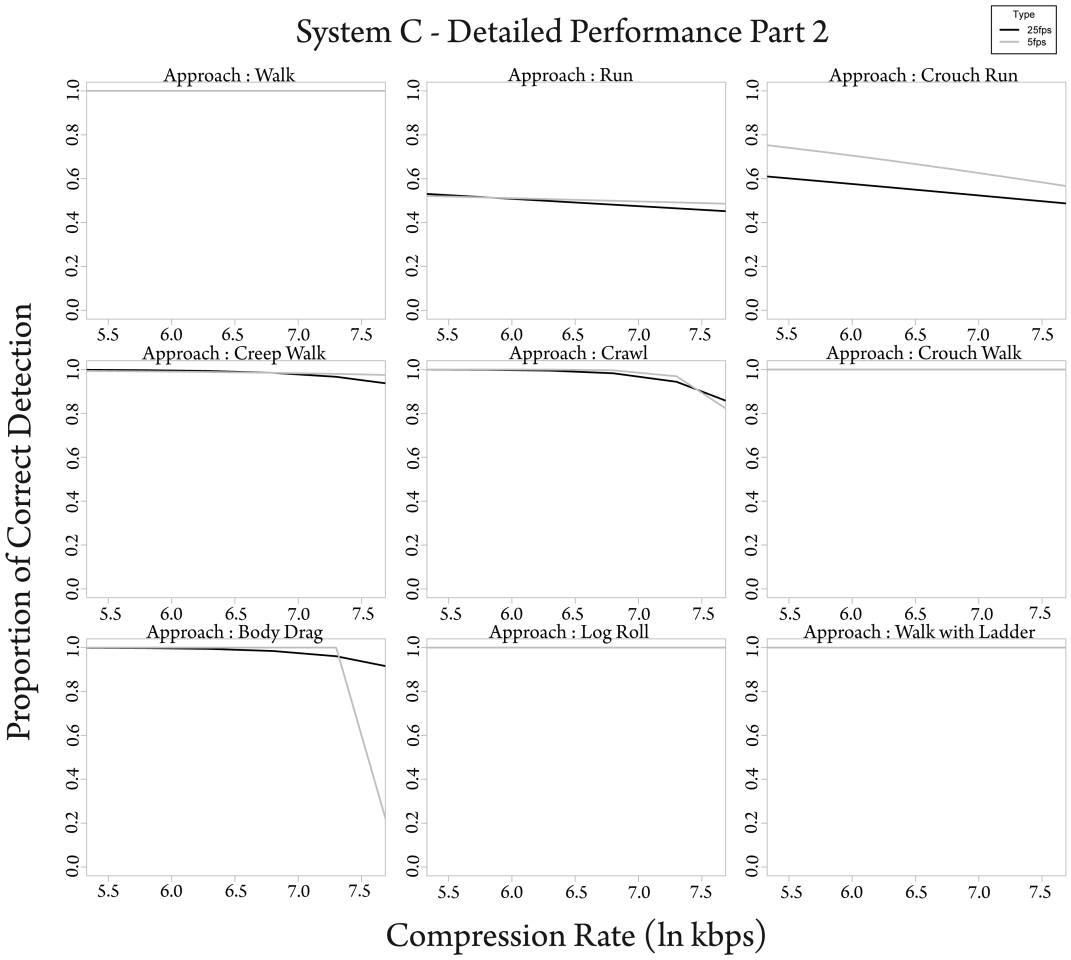


Figure 6.11: Detailed performance with respect to compression (in ln kbps) for system C Part 2 (as graphs in Figure 6.6).

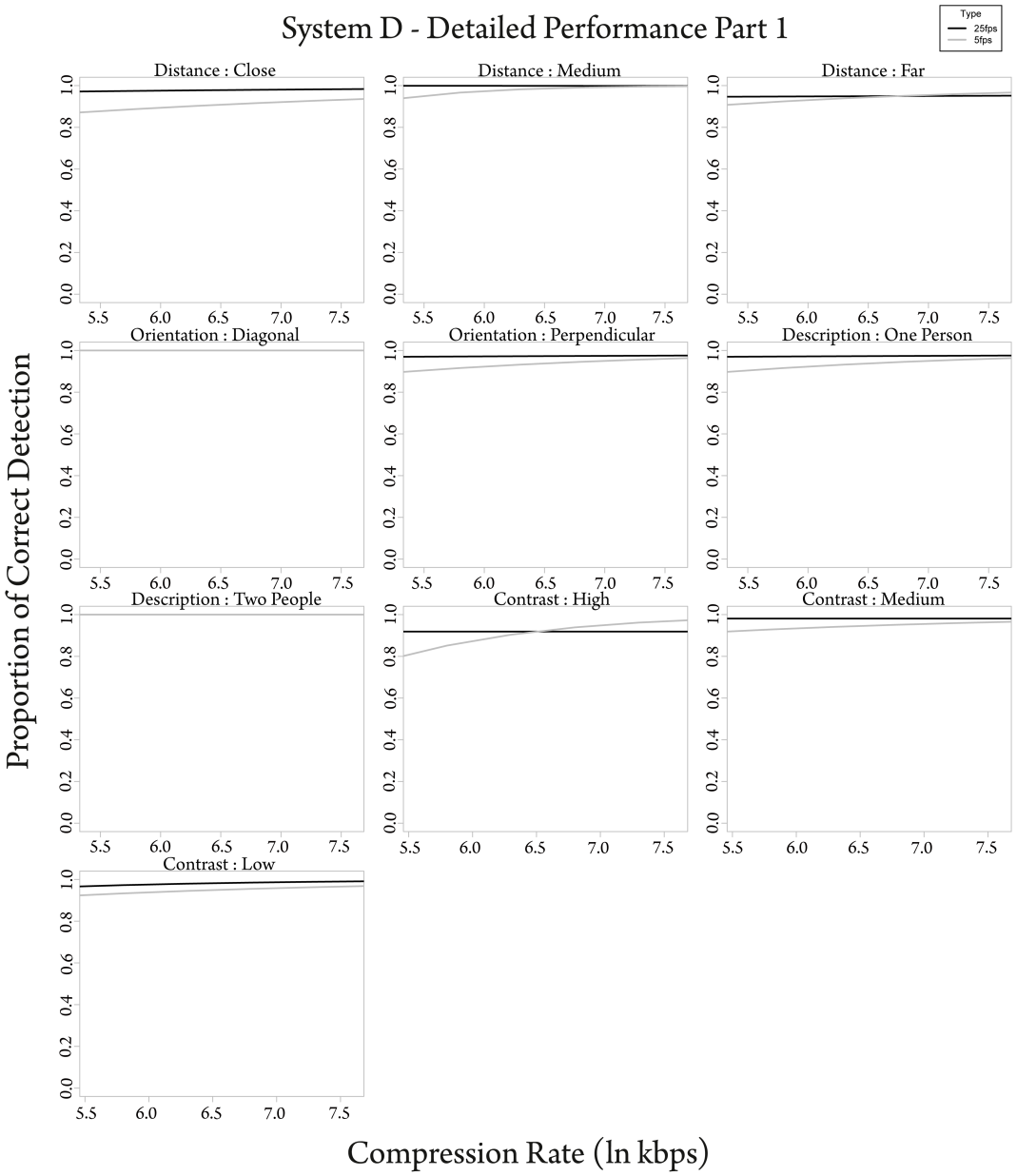


Figure 6.12: Detailed performance with respect to compression (in ln kbps) for system D Part 1 (as graphs in Figure 6.6).

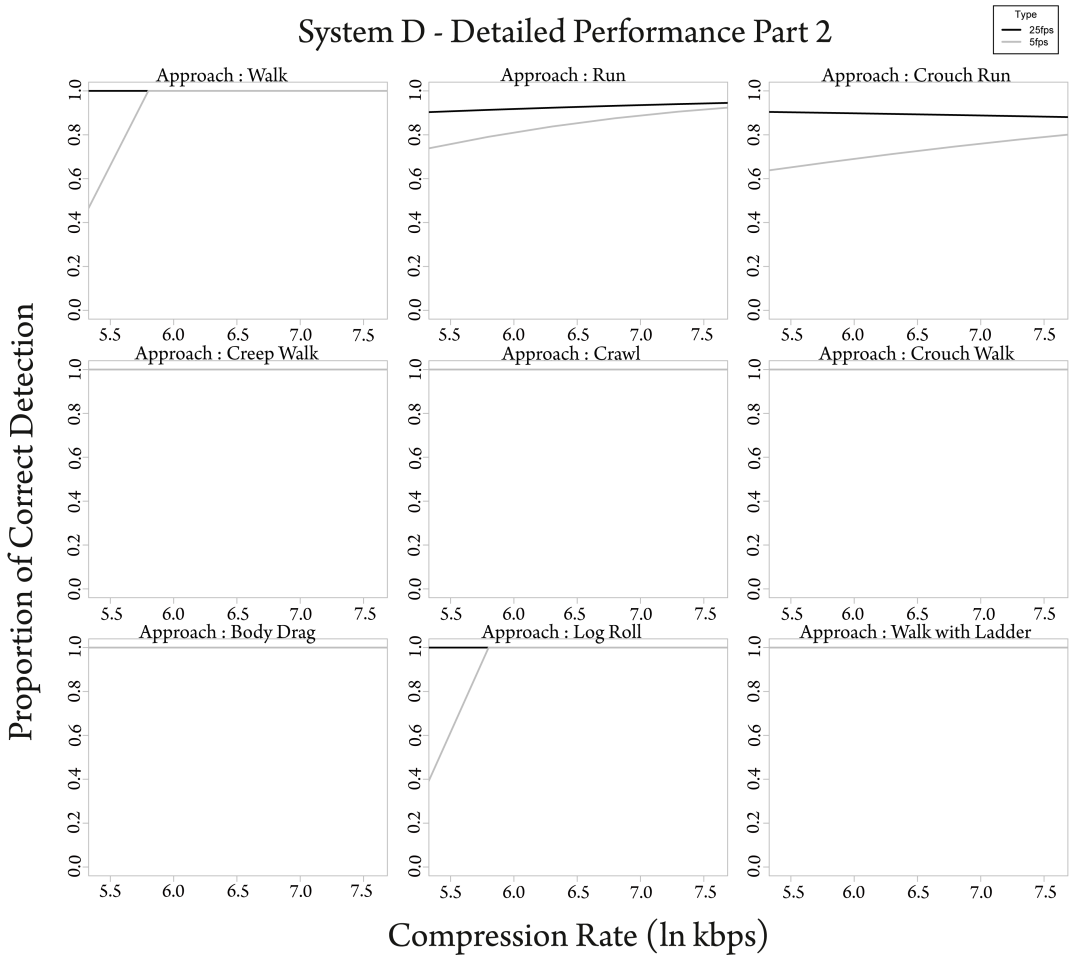


Figure 6.13: Detailed performance with respect to compression (in ln kbps) for system D Part 2 (as graphs in Figure 6.6).

Observing the obtained logistic regression models for System A (see Figures 6.6 and 6.6), one can identify the scene properties that result in a decline in the performance of system A. These are scenes that exhibit properties such as close distance, high and low contrast, run and crouch run approach. Furthermore, footage of 5fps type has been affected more than 25fps. This is expected as often analytics systems incorporate tracking techniques and their performance is dependent on motion continuity (see Section 2.1.3). Additionally, at very high compression amounts (200kbps at 25fps and 40kbps at 5fps) performance declines for half of the scene properties under investigation (e.g. medium and far distance, diagonal orientation, crawl, crouch walk, body drag and walk with ladder approach). These results (logistic regression) agree with the results from linear regression in Figures C.2 and C.3 (see Appendix C).

System B has provided different results; the scene properties that have reduced the performance of system B are close distance, high contrast, run and crouch run approach (see Figures 6.8 and 6.9). Reduction of performance is more obvious with scene properties such as crawl, body drag and log roll approach. Compression has increased performance for the high contrast property (at 25fps) and the walk approach (at 5fps). Overall, footage of 5fps type has been affected more than 25fps. Additionally, these results agree with the linear regression analysis in Figures C.4 and C.5 (see Appendix C).

System C has performed the worst with scene properties of run and crouch run approach (see Figure 6.11). In contrast, scene properties such as close distance, creep walk, craw, and body drag approach properties have benefited from compression (see Figures 6.10 and 6.11); increase of compression amount has resulted in the increase of system performance. In most cases, 5fps type of footage has derived the same results as the 25fps; expect for the crouch run property were better performance has been obtained with the 5fps footage. Similar detailed results to the ones obtained from logistic regression have been derived from the linear regression analysis in Figures C.6 and C.7 (see Appendix C).

Lastly, system D has produced the worst performance for scene properties: run and crouch run approach (5fps worst than 25fps), close distance at 5fps, high contrast at 5fps, and walk and log roll approach at 5fps and at 40kbps (see Figures 6.12 and 6.13). These results are similar to the ones obtained from the linear regression in Figures C.8 and C.9 (see Appendix C).

In conclusion, the most common scene properties that affect the performance of the video analytics systems under test are: close distance (for systems A, B, C and D), high contrast (for systems A, B and D), run and crouch run approach (for systems A, B, C and D), body drag approach (for systems B and C), and log roll approach (for systems B and D). Some of the close distance scenes have decreased the performance of VA systems simply because the subject/attacker in the scenes might have been confused for a spider on the lens so the systems have not performed detection.

Sys.A	**kbps at 25fps**							**kbps at 5fps**					
	Ref	2000	1600	1200	800	400	200	400	320	240	160	80	40
Clip 1	none	none	none	none	none	none	2	none	none	none	none	none	5
Clip 5	14	10	9	10	none	11	27	8	8	2	9	4	19
Clip 7	none	none	none	none	none	none	none	2	none	none	none	none	10
Clip 12	246	210	209	193	229	198	209	198	194	119	195	192	212
Clip 15	none	none	none	none	none	none	none	none	none	none	none	none	45
Sys.B	**kbps at 25fps**							**kbps at 5fps**					
	Ref	2000	1600	1200	800	400	200	400	320	240	160	80	40
Clip 5	none	none	none	none	none	none	none	none	none	none	none	none	1
Clip 10	none	none	none	none	none	none	none	none	none	none	none	none	2
Clip 12	2	4	5	2	6	9	77	3	9	9	11	10	77
Sys.C	**kbps at 25fps**							**kbps at 5fps**					
	Ref	2000	1600	1200	800	400	200	400	320	240	160	80	40
Clip 1	none	none	none	none	none	none	2	none	none	none	none	none	6
Clip 2	none	none	none	none	none	none	none	none	none	none	none	none	5
Clip 5	none	none	none	none	none	none	8	none	none	none	none	none	28
Clip 6	none	none	none	none	none	none	8	none	none	none	none	none	4
Clip 9	none	none	none	none	none	none	none	none	none	none	none	none	5
Clip 10	none	none	none	none	none	none	4	none	none	none	none	none	3
Clip 12	none	none	none	none	none	none	229	none	none	none	none	none	280
Clip 13	none	none	none	none	none	none	none	1	none	none	none	none	none
Clip 14	none	none	none	none	none	none	1	1	none	none	none	none	7
Clip 15	none	none	none	none	none	none	1	2	none	none	none	none	1
Sys.D	**kbps at 25fps**							**kbps at 5fps**					
	Ref	2000	1600	1200	800	400	200	400	320	240	160	80	40
Clip 12	none	none	none	none	none	none	1	none	none	none	none	none	4

Figure 6.14: Total number of false alarms for each VA system for the 10 times repeated trials.

Spiders are under the distraction category and are very common on CCTV lenses of outdoors systems. Furthermore, each system has performed differently for each scene property and universal conclusions on what constitutes good image quality for automated tasks cannot be reached. Instead, designers and testers of automated systems can identify the image acceptance of specific scene properties for their system.

6.3.3 False alarms

Figure 6.14 consists of 4 sub-tables that correspond to each of the 4 VA systems. The sub-tables provide information on the system name, clip number (clip description can be found in Table 6.1), amount of compression and number of frame rates (25fps or 5fps), and the total number of false alarms that occurred from the 10 repeated trials. For example, system A (Sys. A) produced 210 false alarms (i.e. an average of 21 false alarms 210/10) with compressed footage at 2000kbps and 25fps for clip 12. 'none' indicates zero production of false alarms. Some compression levels are missing in the sub-tables for systems C and D as no false alarms were produced for these missing compression levels.

Most false alarms (see Figure 6.14) were triggered with distraction clip 12, which was filmed on a sunny day. Clip 12 contains small clouds in the sky causing many abrupt illumination changes and moving shadows through the fence (see Table 6.1 for clip description). Not many false alarms were produced from the clips containing attacks.

6.4 Discussion

This chapter has provided a methodology for testing automated algorithms with uncompressed and degraded footage. The results have shown that the proportion of correct attack detections for systems A and D at 5fps increases significantly with increasing bitrate (less compression). For the rest of the compression levels and

systems, compression has not affected the overall performance of the systems. An analysis based on the scene content properties enables analytical understanding on where systems need improvement. Each system, depending on how it has been designed, was shown to be affected negatively or positively by the scene properties under investigation.

The specifics of the systems under test are unknown and universal conclusions cannot be made. For example, other systems might produce totally different results. But, adopting a methodology similar to the one employed in this chapter will allow manufacturers to identify the scene properties for which their system might need further improvement. For example, if a system does not perform optimally (100% correct detection) with a close distance scene property then additional footage with that specific property (in combination with other properties for example colour of clothing, run approach) can be utilised/experimented to improve performance.

The findings in this investigation do not agree with the subjective results reported in Chapter 4. For example, for the camera to subject distance property the far scenes produced lower subjective scores than the close scenes (closer distance scenes provide more visual information for humans). In some VA systems, the close distance attacks produced lower scores in comparison to far, and medium distance attacks. This confirms that the term image quality should not be used in the same manner for automated and human visual systems. Instead image acceptability in terms of scene content property acceptance to complete the recognition/identification task is a better-suited definition for automated systems. The criteria/thresholds of automated human detection systems do not necessarily match those of humans. This is promising in terms of having automated systems to do tasks that humans find difficult. For example, most automated systems were affected more from the close distance image property than the far distance. Overall, the automated systems (only the ones tested in this thesis) can do better for the far distance scenes than humans and humans can do better for the close distance scenes. Both humans and automated systems could benefit from each other. For example, automated systems

can benefit from using human perception, in this case, perhaps for object/subject description.

Additionally, it is important at the collection stage of any dataset to make sure that scene content properties are taken into consideration to enable the collection of varied scene contents and degrees of difficulty. For example, in this investigation most false alarms were triggered with a clip that contained many abrupt illumination changes. Further, not much variability in terms of distractions can be noticed in Table 6.1 for system designers to experiment with. Also, in Table 6.2 the 11 hours dataset under test includes only 3 scenes with the log roll approach property. This type of approach might be considered the most likely one for intruders.

CHAPTER 7

Discussion

This chapter contains a critical discussion on the findings obtained from the experimental part of the thesis (Chapters 4, 5 and 6). The critical discussion identifies potentials, weaknesses and limitations of the employed methodologies and obtained results. The findings are discussed in connection with the subject as outlined in the background (Chapters 2 and 3) in order to define what new has been achieved in this work.

The main aim of this research is to investigate aspects of image quality and video compression that may affect the completion of police visual tasks with respect to CCTV imagery. CCTV systems commonly operate in semi-controlled (e.g. bus CCTV systems, door surveillance) or uncontrolled (e.g. open street CCTV systems) environments and their produced imagery is normally compressed in order to compensate for storage or transmission requirements. In summary, 7 factors have been identified that will influence image usefulness (or amount of useful available visual information) of CCTV systems and further the completion of a recognition task.

These 7 factors are: 1) subject to camera distance (e.g. a close distance may allow facial recognition and a further away distance, gait or clothing recognition), 2) angle of camera to the subject/object (e.g. frontal face view entails more information than a tilted face view), 3) illumination conditions (i.e. intensity, colour, angle of illumination - result in the production of over, under, mixed and correctly exposed scenes), 4) system performance (i.e. sensor, lens, image processing), 5) recording/transmission (i.e. spatial and temporal compression), 6) the fitting of wrong cameras (i.e. if the location is wrong for the task or if the cameras are not working) [2], and 7) occlusions (e.g. for a face recognition task that could be hats/glasses and for an automated human detection task that might be cars passing in front of the human to be detected). These 7 factors represent the variability present in real-world CCTV applications and they need to be taken into consideration as they affect the amount of captured visual information. In order to limit the effect of the 7 factors on the captured information, they need to be taken into consideration. For example, to apply compression upto an acceptable level or the installation of CCTV systems according to specific task requirements (e.g. face recognition tasks require the camera to be closer to the subject in comparison to the people counting tasks).

Image usefulness or image appropriateness for the completion of police tasks depends on the specifics of the task. For example, the imaged information between

faces and clothing differs and different compression amounts would be acceptable for each type of task. This is the reason why specific tasks are investigated in this thesis. Additionally, the term image usefulness is used in the same manner between automated and human visual systems. This thesis has included both human and automated visual systems to identify if there is a correlation between them concerning image usefulness and video compression.

In summary, 2 types of visual systems (human and automated) for 3 specific police tasks (human face recognition, automated face recognition and automated human detection), are assessed with characterised CCTV imagery and video compression (264/MPEG-4 AVC). The CCTV footage has been characterised in terms of defined scene content properties as the performance of imaging systems/processes (e.g. subjective investigations, compression algorithms, automated recognition/detection systems or human face recognition) is dependent on scene content (see Chapters 2 and 3). Furthermore, knowing exactly with what content characteristics a system (e.g. automated systems, compression algorithms) fails can contribute to the further improvement of such a system. For instance, if an automated system does not perform well with under exposed scenes then manufacturers can develop techniques concentrating in improving performance with under exposed scenes. Another example is to utilise compression up to an acceptable level. The following paragraphs concentrate on developing a discussion around the objectives of this thesis as stated in Section 1.1.

Information relating to the datasets utilised in the investigations was provided in Sections 4.2.1 (face dataset) and 6.1 (sterile zone dataset). The main aim, when developing a dataset, is to cover scene content information/variability that is commonly encountered by the visual system under evaluation. The created CASTBUS 2012 dataset has succeeded in covering scene content properties, encountered by a challenging (in terms of illumination), real-world application: the London bus. The illumination variations were obtained because the filming took place on a sunny day. This would not have been the case on an overcast day: the clouds would have

produced diffused and uniform illumination, which would have resulted in the production of correctly exposed scenes. The CASTBUS 2012 dataset provides content variability in terms of: camera to subject distance, spatial-temporal busyness, illumination conditions and facial angles to the camera plane. CASTBUS 2012 has allowed the capture of varied scene content properties.

The main objective of this thesis is to identify the scene properties that influence police tasks and CASTBUS 2012 has contributed towards that objective. Scene content characterisation allows the testing of systems with known properties. This perhaps can be understood as being related to the process employed to assess a system's tone transfer function. Transfer functions define a relationship between input signal (i.e. densities of a test target, pixel values displayed on a monitor) and the system's output (i.e. pixel values of a camera system, luminance of an LCD system) response to that signal. This is where this investigation is different from the rest, the input signal has been specified in detail by including scene content characterisation and the outputted signal has been analysed taking into consideration that input signal (i.e. individual scene properties). No other investigation, relating to human or automated face recognition or automated human detection, has been identified that takes into consideration scene content characterisation to the same extent as the research included in this thesis.

There is another dataset available, representative of a real-world CCTV application, called SCface (surveillance cameras face database) [294]. This dataset includes footage from the entrance of a doorway utilising 4 indoor CCTV cameras. The SCface dataset does not include much variability in scene content. For instance, only 1 facial angle is included. There are numerous other facial datasets [295, 296] but all of them seem to have certain limitations. Some of them include only still imagery (not appropriate for assessing video compression), most are captured under semi-controlled or controlled conditions, and some others are in high definition format. Most CCTV systems operate in standard definition format.

In case of the automated human detection task, an already available dataset was

utilised the Sterile Zone (SZ) scenario from the iLIDS scenarios [49,50]. The manufacturers of the 4 video analytics systems under investigation have developed their systems based on this specific SZ scenario and there was no option of using another scenario. Employing in the investigation an existing dataset might raise some concerns in terms of the appropriateness of that dataset for the investigation. In Chapter 6, the effect of scene content properties on the performance of video analytics were investigated. The utilised SZ dataset does not include equal (or even inadequate) numbers of scenes for each individual property (see Table 6.2), which creates challenges in assessing the performance of the VA systems under investigation. For instance, the properties walk and log roll from the attack approach category include 28 and 3 scenes respectively. Most systems have always detected the log roll approach property apart of System B (see Section 6.3.2); the number of scenes for the log roll property can be considered small in comparison to the walk property approach. The content properties that include fewer than 10 grouped scenes are: crouch walk, body drag, log roll and walk with ladder from the approach category, and the high contrast from the contrast category. This corresponds to 5 out of the 19 scene properties under investigation consisting of fewer than 10 grouped scenes.

Also, the number/type of distractions that are included in the utilised SZ dataset (see Table 6.1) can be considered small. For example, outdoor CCTV systems often attract spiders on the lens and no such a distraction was present in the dataset. As mentioned in Section 6.2.1, the utilised dataset was selected based on the availability of the original tape recording (i.e. DigiBeta videocassettes) of the SZ scenario in order to have the footage in an uncompressed format. The other option would have been to have more footage (i.e. explore more scene properties and distractions) but in a lightly compressed format (i.e. 10% compression). This option was avoided as it was expected that the VA systems might increase performance in comparison to the 'uncompressed' reference with small amounts of compression. This has been proven true as illustrated by the graphs of the Yes scenes in Figures 6.3 and 6.4, where some compression amounts have increased performance in comparison to the

reference 'uncompressed'. As this has been established, in a future investigation the 10% compressed footage that includes additional scene content properties could be utilised.

No information has been found in relation to how many scenes should be included in each grouped property in order to derive statistical significant/valid results. For example, defining the number of scenes adequate for a statistical analysis with the low skin/face lightness property. Most statistical methods refer to sample sizes of people. Nevertheless, some observations can be made from the models fitted to grouped properties (e.g. close camera to subject distance property) in Chapters 5 and 6. The human investigation in Chapter 4 has fitted models to individual scenes and not to grouped properties.

In the face recognition investigation (i.e. automated and human) in Chapter 5, as the number of grouped properties under each category increase, it became more difficult to derive conclusions on the derived results. For example, the skin lightness category consists of 5 types of lightness properties (i.e. low, high, medium and mixed lightness groups), which are distributed across 25 scenes. A greater number of scenes for each grouped property would had decreased the derived errors on the models, as the reliability to reflect the population mean increases with having a larger sample, making the comparison among the models statistically easier. Nonetheless, the derived models for the lightness group do illustrate tendencies and should not be considered inadequate. For example, all 3 of the automated face recognition systems studied have performed the best with compressed 'low lightness' scenes and the worst with compressed 'mixed lightness' scenes. Also, 2 different statistical approaches have been employed for the human data and the same conclusions were drawn (i.e. investigations in Chapters 4 and 5).

In the human detection task in Section 6.3.2, the proportion of correct detections for each scene property (e.g. walk approach) is plotted against compression rates. Perhaps a more balanced number of scenes/attacks should have been included under each property (see Table 6.2). This does not make the derived results invalid,

since the 2 applied statistical methods have produced the same results (see Section 6.3.2 for the logistic regression analysis and Appendix C for the linear regression analysis).

Overall, representative CCTV footage has been produced both 'uncompressed' and degraded (i.e. compressed and with reduced frame rate), utilising 2 implementation types of H.264/MPEG-4 AVC (i.e. MPEG Streamclip for both face recognition and human detection tasks, and CCTV DVRs only for the face recognition task), and with different scene content properties. This material can be released to manufacturers of CCTV systems in order to enable further testing of their systems.

A couple of different camera systems were utilised between the face recognition (Chapters 4 and 5) and human detection (Chapter 6) investigations. A standard definition CCTV camera system was utilised for the human detection investigation. No further information is available in relation to this CCTV camera system as an already standard available dataset has been employed (i.e. the SZ scenario part of the iLIDS datasets). There are numerous companies that provide CCTV systems and their quality has never been studied or quantified. Also, perhaps for the human detection task the quality of the camera system (unless it has been extremely degraded) might be of a lesser importance in relation to other factors (e.g. way of approaching the fence such as run or log roll). Furthermore, the detection of a human silhouette task would not require as much image content information or detail in consideration of a face recognition task. Faces include finer details (i.e. shape and distance of individual features), where the capture of these details would contribute to the completion of a face recognition task. Otherwise, all faces include a nose, mouth and eyes.

In the human and automated face recognition tasks, instead of a CCTV camera system a consumer-quality DV camcorder was utilised for various reasons including quality, accessibility and cost. The consumer DV camcorder has produced overall higher image quality output than the typical sample CCTV camera installed on London buses (see Figure 4.1). There are pros and cons in utilising either camera

system (CCTV camera system or DV camcorder) in face recognition investigations. For instance, as there are numerous companies that provide CCTV systems with varied and unknown qualities, perhaps the DV camcorder provided a starting 'standardised' quality. An option to compensate for the quality differences between the 2 camera systems is to apply a frequency filter aiming to visually match the frequency response of the DV camcorder to that of the CCTV system. Furthermore, as technology advances CCTV systems will produce in the future comparable image quality to that of consumer video systems. On the contrary, it is unknown if the results, from the 2 face recognition investigations would have been different if a CCTV camera system was utilised instead of the DV camcorder. Yet, perhaps again the quality of the camera system might be of a lesser importance in relation to other factor such as illumination conditions. For example, under the same variable illumination conditions both DV camcorder and CCTV systems are expected to produce under, over, correctly and mixed exposed scenes.

In the human face recognition investigation in Chapter 4, the applied methodology was aimed to obtain acceptable compression limits for the London bus application. First an industry implementation (MPEG Streamclip) of the standard H.264/MPEG-4 AVC was utilised to identify from a selected set of 25 scenes the key scenes (i.e. the scenes most affected by compression). Later, the pre-selected key scenes were utilised with 5 of the most commonly used CCTV recording systems on London buses to identify the acceptable compression limits for that application. Observing Table 4.10, the results obtained from the CCTV DVRs and the industry standard compression (i.e. MPEG Streamclip) agree in terms of which scenes required lighter compression (i.e. higher number of kilobits per second). These were the low and high lightness scenes in comparison to the medium (daylight and bus illumination) and mixed lightness scenes. In conclusion, low and high lightness scenes are affected by H.264/MPEG-4 AVC compressor the most for human face recognition tasks, and this is valid for any implementation of H.264/MPEG- AVC (i.e. CCTV DVR proprietary and industry-standard).

Furthermore, in Section 4.4, the unpredictable nature of the CCTV DVR systems is presented when reducing the frame rate from 25fps to 4fps. For example, the reduction of the frame rate has output 1 image from the 8 images (i.e. corresponding to 8 successive frames in the footage) of the face in the reference footage. This outputted 1 image might be the worst, or the best-case scenario from the 8 available images of the face (see Figure 4.10). Additionally, the CCTV DVRs have performed some processing on the recorded/compressed footage in terms of making dark areas appear brighter (made face information more visible) and sharpening of edges. This has affected the results as the average fit for the CCTV DVRs to the data points (see Table 4.10) has outperformed the industry standard compressor. This does not mean that the image itself includes more information than the images produced from the consumer industry standard compressor. For the industry standard compressor, no contrast or tonal processes on the compressed footage have been observed. Also, the tone characteristics of the industry standard compressor are similar to the 'uncompressed' reference confirming that no processes on tonal information have taken place (see Figure 5.14 and Section 5.3.4). Performance of human face recognition tasks could be helped by applying illumination normalisation techniques on face images (e.g. adjusting the levels of a dark scene to appear brighter see Figure 3.9).

Additionally, image sharpening combined with compression may create information in a compressed image that might not be present in its reference 'uncompressed' version. For example, as mentioned in Chapter 1, the brain has the ability to put together information from what it knows combined with the information that is seeing (i.e. the fillin-in phenomenon [8]). When a facial image is over sharpened and compressed at the same time, perhaps 'true' facial information might be reduced due to the blocking artefact and altered (i.e. due to the sharpening of the ringing artefact and/or the edges of blocking artefacts). This might create a visual illusion that facial information is present in the compressed scenes. Whereas, in reality the facial information might have been reduced from compression and/or altered from sharpening of artefacts.

In comparison, the results obtained from the automated systems (see Section 5.3.3) have shown that the systems have performed better with the industry standard compressor than the CCTV DVR compressors. As mentioned in Section 5.2, the CCTV DVRs have altered the size of the original reference scenes (i.e. by recording at lower resolution). For this reason, the selected facial regions between the reference scenes and compressed scenes with the CCTV DVRs are not exactly the same. The selected facial regions between the reference scenes and scenes compressed with the industry standard compressor are exactly the same. The non inclusion of exactly the same facial regions between the compressed scenes with the CCTV DVRs and the reference scenes might be the most important reason for the automated systems perform worse with CCTV DVRs.

The methodologies adopted for the 2 face recognition investigations (human and automated) are similar in terms of both being assessed based on the difference between a degraded image from its reference version. The methodology that was followed for the human investigation had been put together from well-established methods used in psychophysics. This is applicable from the presentation of the stimuli (e.g. when assessing usefulness/distortion the reference is provided), to the collection of observers' responses (e.g. the *yes*/*no* responses and instructions to observers) and analysis of results (i.e. the fitting of psychometric function). Automated face recognition systems are normally assessed by analysing correct recognition rate from large datasets. Also, detailed information on scene contents are normally not included. The advantage of the current methodology employed in Chapter 5 over the 'standard' one, is that it provides a detailed analysis based on scene content properties, and identifies numerically the distance between degraded and reference facial images. Overall, the current employed methodology provides a detailed analysis, whereas the 'standard' one provides a general analysis such as the recognition rate has increased or decreased with compression. Perhaps, the current methodology can be adopted to incorporate aspects of the 'standard' methodology. For instance, to provide results based on correct recognition for individual scene properties. This is also applicable for the methodology employed in the assessment of video analytics

systems. Similarly to automated face recognition systems, video analytics systems are normally assessed by obtaining correct recognition rates from large datasets. In the employed methodology for the video analytics systems in Chapter 6, the results are analysed in detail based on individual scene properties.

The results from the human face recognition task have been analysed utilising 2 different approaches, both of them have drawn the same conclusions. In the first statistical approach individual psychometric curves were fitted to the results obtained from the individual scenes. Later, the points at 75% (in kbps) on the curves were selected for further analysis aiming to identify any significant trends such as similarities and differences between the classified and grouped properties. In the second statistical approach, psychometric curves were fitted to grouped properties; to all the data of the grouped scenes with the same content property. The performance of human face recognition tasks when assessed with compression is affected the most negatively by scenes exhibiting 'low lightness', 'far camera to subject distance' and 'high spatial-high temporal busyness' properties. In comparison, positive performance (or least degraded performance) with compression has been obtained with scenes exhibiting 'medium lightness' (i.e. for both bus and daylight illuminations), 'close camera to subject distance' and 'low spatial-low temporal busyness' properties.

The data derived from the automated face recognition systems have been analysed/modelled in a similar statistical approach to that of the human face recognition data in terms of fitting curves (i.e. in this case regression models) to scenes with the same grouped property. The performance of automated face recognition tasks when assessed with compression is affected the most negatively by scenes exhibiting the 'mixed lightness' scene property. In case of the LDA method, 'tilted' angle to the camera plane scenes have produced lower scores than 'frontal' scenes. The least degraded performance with compression, has been obtained with scenes exhibiting 'low lightness', and 'low spatial-low temporal busyness' properties. The results from the latter property are the only ones that agree with the results obtained from the

human data, indicating no correlation of image usefulness between automated and human face recognition systems. Predominantly, for all the investigations, humans were affected the most by compression in comparison to the automated face recognition and video analytics systems. Even though compression algorithms are meant to throw away information that humans cannot perceive when it is employed at high amounts then visible information is lost.

Furthermore, the positive performance of the automated face recognition systems with 'low lightness' scenes has been discussed in detail in Sections 5.3.4 and 5.4. As the tone characteristics do not change in the compressed scenes (i.e. for the industry standard encoder) with respect to the 'uncompressed' reference, and the dark areas within a 'low lightness' scene might occupy a large portion of an image (i.e. in comparison to the medium lightness scenes), it is more likely that the automated algorithms performed pattern/lightness matching between the dark areas of the facial images in the 'low lightness' content category. This could be an important reason why the 'low lightness' scenes were affected the least even at high compression levels. On the other hand, the CCTV DVR encoders have altered the tonal characteristics from the reference, making the dark areas appear brighter, and have produced similar results to the industry standard encoder in terms of the 'low lightness' scenes to have been affected the least by compression. It is unknown how much the tonal characteristics of the reference have been altered by the CCTV DVRs. For example, the compressed 'low lightness' scenes with the CCTV DVR encoders still appear darker than the 'medium lightness' scenes. The same methodology should be applied after processing the facial images with an illumination normalisation technique in order to identify if the 'low lightness' scenes are still affected the least by compression. Perhaps, 'low lightness' scenes entail different information from 'medium lightness' scenes concerning image spatial frequencies. The results obtained from this investigation are only valid for the systems under test; other systems or other versions of the utilised systems might produce different results. The employed methodology could be used to assess any automated face recognition system.

In Chapter 6, 4 video analytics systems were assessed with: compression, 2 types of frame rates (25fps and 5fps), scene content properties and distractions. The results in Figures 6.3 and 6.4 have shown that every system performed differently for each compression/frame rate level (see the Yes, No and Uncertain scene graphs), but overall compression has not adversely affected the performance of the systems. The reduction of frame rate from 25fps to 5fps has decreased detection performance for most systems. This is understandable as analytics systems utilise the continuity of a video content to perform analysis/detection and that continuity is disrupted when reducing the frame rate to 5fps. The detailed analysis on the performance for each system has identified the most common scene properties that cause a decline in the performance of the video analytics systems under test. These are: 'close distance', 'high contrast', 'run approach', 'crouch run approach', 'body drag approach' and 'log roll approach'. If this were a human investigation then one might expect the 'close distance' and 'high contrast' properties to have contributed to the increase of detection performance. The investigation in this thesis proves that this is not the case for the automated analytics systems. The results obtained from this investigation are only valid for the video analytics systems under test and other systems might perform differently. These systems have received UK Government approval and are considered operationally successful (see Section 6.1).

The selected scene properties that have been included in the 3 investigations might be inadequate, or additional properties might be required, to describe the content of images for police tasks. In the human investigation in Chapter 4, the results could have been influenced by the degree of distinctiveness or overall appearance of the actual faces in the scenes such as a big nose, head and eyes (see Section 4.3.1). For instance, a face with big facial features will require less compression than a face with smaller facial features. This is applicable in particular for human face recognition tasks where subject to camera distance has been identified as an influential factor that affects performance of these tasks (see Table 4.8). In addition more groups of the existing properties could have been included such as greater degrees of facial angles to the camera plane or further/closer camera to subject distance groups.

In the case of the video analytics systems, the intruders in the footage were wearing only 2 types of clothing: a) white, could be considered to produce high contrast properties with respect to the green grass background, or b) green, could be considered to produce low contrast properties with respect to the green grass background. The latter is acting as a camouflage (see Section 6.2.2). It is unknown what would happen in a real case, where the intruder might wear another colour. Perhaps, the inclusion of more colours of clothing in the dataset should be considered.

CHAPTER 8

Conclusions and further work

The conclusions from this work are as follows.

Humans were affected the most by compression in comparison to the automated face recognition and video analytics systems. This indicates that humans are more sensitive to the removal, by compression, of visible information than automated systems. Even though compression algorithms are designed to through away information that humans cannot see, when employed at high compression levels (as is the case in the security industry) it does through away visible information. This is not the case for automated systems and footage can be compressed at high levels and still not affect their performance. Perhaps automated systems utilise different image content information to complete a task than humans. This can only apply to those systems/algorithms studied. Additionally, the performance of the automated systems with compression is a positive output in terms of saving storage for automated tasks. Yet, the end users of security imagery are the police officers (e.g. they arrest suspects) and the court (e.g. they establish identities) as they are the executors of the justice system.

Overall individual scene properties need to be taken into consideration when assessing visual systems for police tasks. The performance of human face recognition tasks when assessed with compression is affected the most negatively by scenes exhibiting 'low lightness', 'far camera to subject distance' and 'high spatial-high temporal busyness' properties. The least degraded performance has been obtained with scenes exhibiting 'medium lightness' (i.e. for both bus and daylight illuminations), 'close camera to subject distance' and 'low spatial-low temporal busyness' properties.

In comparison, the performance of automated face recognition tasks when assessed with compression is affected the most negatively by scenes exhibiting the 'mixed lightness' scene property. In the case of the LDA method, 'tilted angle' to the camera plane scenes have produced lower scores than 'frontal angle' scenes. The least degraded performance has been obtained with scenes exhibiting 'low lightness', and 'low spatial-low temporal busyness' properties. The results from the latter property are the only ones that agree with the results obtained from the human data; indicating no correlation of image property acceptance between automated and human face recognition systems. This is not necessarily good or bad, humans and automated systems are just different. This can be good in terms of having more tools to solve a recognition task. In most cases both humans and automated systems are combined in police tasks. Understanding the scene dependency phenomenon helps in incorporating within testing methodologies various scene content properties that are representative for the task. Testing methodologies are used in order to understand how systems behave and where they need improvement (e.g. over exposed scenes for the automated face recognition algorithms: LDA, KFA and PCA). More research is required to identify if scene dependency is a phenomenon that needs to be taken into consideration for any automated system as the current investigation includes a limited number of systems.

The most common scene properties that reduce the performance of the video analytics systems under test are: 'close distance', 'high contrast', 'run approach', 'crouch

run approach', 'body drag approach', and 'log roll approach'. If this were a human investigation then one might expect the 'close distance' and 'high contrast' properties to have contributed to the increase of detection performance. The current investigation in this thesis, provides an indication that this is not the case for the automated analytics systems.

The results obtained from the investigations with the automated systems (face recognition, human detection) are only valid for the automated systems/algorithms under test; other systems/algorithms or other versions of the utilised systems/algorithms might produce different results. Furthermore, for the human investigation the results are dependent on the type of observers. For example, civilians might produce different results from police officers.

Furthermore, the employed and developed methodologies in this thesis could be used to assess any visual system aiming to complete a police task.

Findings of this and future investigations could be employed in the creation of quality metrics. For example, a study by Maalouf et al [297] has focused on monitoring quality of legal evidence images in video-surveillance applications by using a combination of a tracking algorithm, a quality metric and a super-resolution algorithm. Furthermore, a more challenging task will be to define quantitatively the relationship between video parameters (e.g. frame rate, bitrate) and image properties (e.g. busyness, lightness) with the acceptability of usefulness of the face for automated and human visual systems. These will need to be different for automated and human visual systems as the results obtained from the included investigations suggest that no correlation exists among them concerning scene property acceptance. The same conclusion has been drawn by Korshunov and Ooi [146] as they have identified that surveillance automated systems (face detection, recognition and tracking) accept signifacantly more compression compared to humans and suggested the need for alternative image quality measures suitable for automated systems.

Future work will involve effort in understanding further the scene dependency

essence of automated face recognition systems (e.g. why under exposed scenes were affected the least by compression for AFR systems). Also, to include illumination normalisation techniques in the evaluation methodology. This additional image processing step might produce different results in relation to image acceptance of scene content properties.

Future work in relation to the video analytics investigation will include the same methodology to be applied on a different more complicated scenario (e.g. traffic monitoring) in order to expand understanding of the performance of automated algorithms.

Appendices

APPENDIX A

Display Characterisation

The EIZO ColorEdge CG210 21.3" Liquid Crystal Display (LCD) was employed for the human investigation in Chapter 4. This investigation was carried out in parallel with another psychophysical investigation by Dr Jae Park. Dr Jae Park has provided information relating to tone reproduction, spatial uniformity, temporal stability and viewing angle characteristics of the employed LCD. This information can be found in his PhD thesis [298]. It was considered suitable, in this case, to use another investigator's data, as both the investigations were conducted at the same time and using the same display settings and viewing conditions.

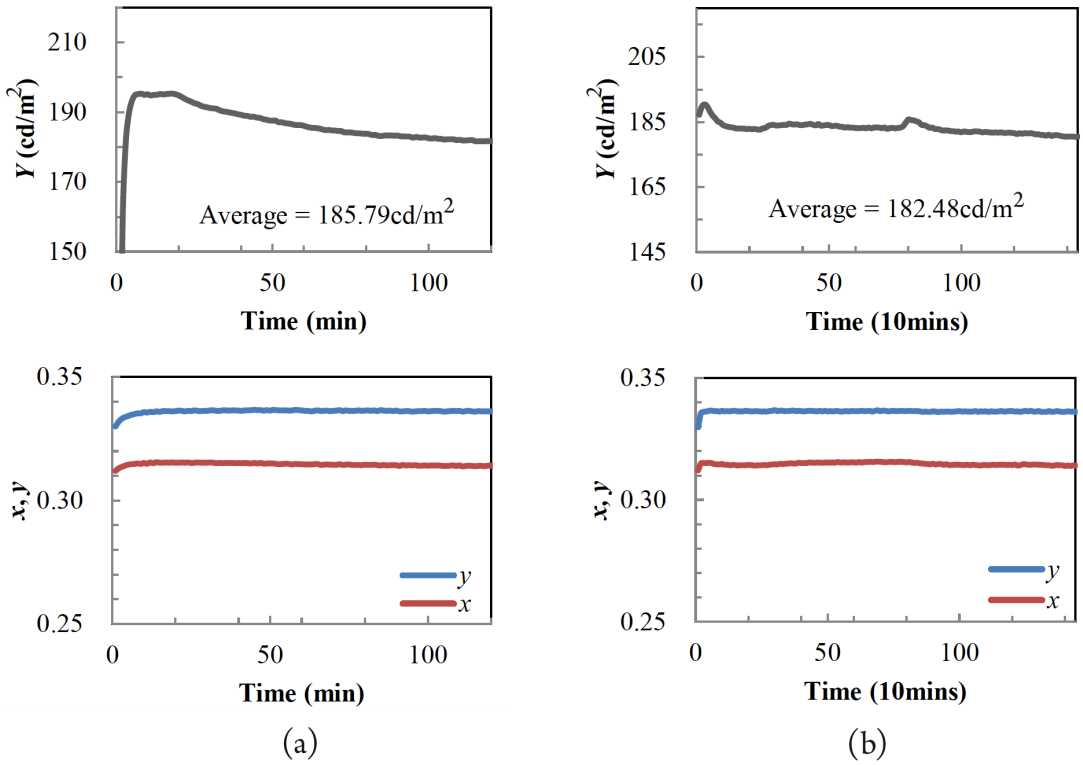
The LCD characterisation was carried out according to the BS EN 61966-4 standard [299]. First, temporal stability measurements were carried out, where the LCD was previously cooled down for 2 hours. The room temperature was constant and approximately 20 (+/-3) degrees Celsius during the performance measurements and psychophysical investigations. Before starting the charaterisation and calibration process, the display was allowed, a warm up time for at least an hour (based on the result from the stability measurement), and as specified in the BS EN 61966-4 stan-

dard. The following points report the calibration and characterisation (performance measures):

- *Daily LCD calibration.* The LCD was calibrated to a white point D65 (6500K), at a luminance of $120cd/m^2$ using an sRGB ICC profile. For the calibration a GretagMacbeth Eye-One Pro system was employed. Although the sRGB standard specifies a white point luminance of 80 cd/m^2 the chosen white point luminance of 120 cd/m^2 is not an uncommon setting in modern LCDs that have generally higher luminance [300]. The result of a higher than the specified white point may produce less accurate display colorimetry, but would not affect the results of the psychophysical experiments, where a standard and a distorted image are compared with each other simultaneously on the same LCD.

- *Temporal stability.* Temporal stability measurements identify colour reproduction instabilities when first applying power (short-term instability) to the LCD and in daily use (mid-term instability) [299]. The LCD device was adjusted to display a white patch and was calibrated to produce a peak luminance of $120cd/m^2$ then it was turned off for a day (the device reset the calibration settings to its original settings automatically when turned off). When the LCD was turned on short-term instability measurements were performed after 1 minute and for the duration of 2 hours. The same procedure was repeated for the mid-term stability measurements, but this time the measurements were performed after 10 minutes and for the duration of 24 hours.

  The measured luminance Y (in cd/m^2) and chrominance coordinates (x, y) are plotted against time (see Figure A.1). Figure A.1a (top) shows that the LCD produced more stable luminance results after only 1 hour warm up time. On the other hand, the chromaticities are stabilised quickly for both short- and mid-term measurements. Figure A.1b (top) indicates a mid-term luminance fluctuation at 80 minutes after switch on, which could be a measurement error, given the otherwise flat luminance response.



(a) (b)

Figure A.1: Temporal stability measurements of the EIZO CG210 LCD. The measured luminance Y (in cd/m^2) and chromaticity coordinates (x, y) are plotted against time for the short-term (a) and mid-term (b) results. Where average is the mean statistical value. Adopted from Park (2014) [298].

- *Tone reproduction.* In this characterisation measurement, the output luminance is plotted as a function of input intensities (i.e. pixel values). The resulting function is commonly referred to as the display transfer function. See Section 3.3.3 on tone reproduction for further information. Figure A.2 provides the transfer functions of the EIZO ColorEdge CG210, where the normalised luminance (i.e. or "normalised light output" on Figure A.2) is plotted against normalised input pixel values for the red, green, blue and neutral display responses [301]. Note that, the tone reproduction (transfer function) is related to the *Y'* (luminance) curves. The measurement involved a total of a 32-step grey-scale ramp, given to the monitor as an input. The Konica-Minolta CS-200 tele-chroma meter (designed for LCDs) was placed 150cm away from the centre and in a parallel plane to that of the monitor. This instrument allowed the measurement of the luminance of the displayed ramp (i.e. obtained by the mean value from 3 measurements for each of the

32 steps), with a measurement angle between 0.1 - 0.2 degrees.

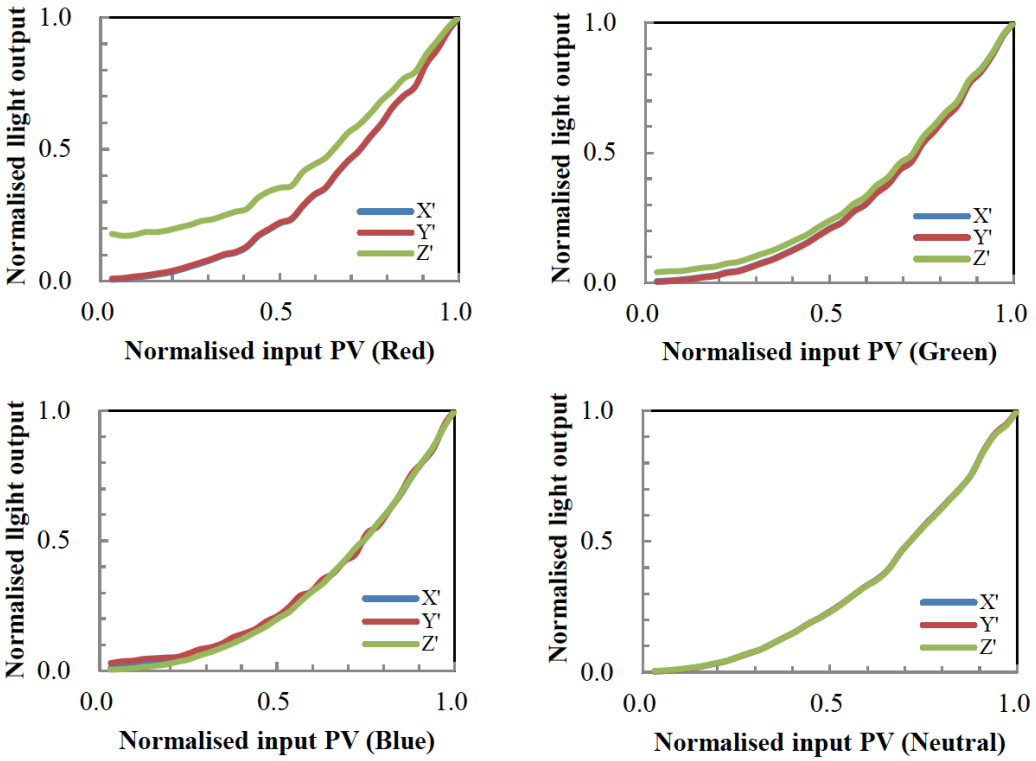


Figure A.2: Tone characteristics of the EIZO CG210 display. Adopted from Park (2014) [298].

- *LCD spatial uniformity.* This involved the display of a white patch (R = G = B = 255) on the entire LCD screen and CIELAB L^*, a^*, b^* measurements of 25 positions across the LCD screen (see Figure A.3). Refer to Section 3.3.4 for information on the CIELAB L^*, a^*, b^* colour space. The variations in lightness ΔL^* was found to be max 6.12, in chroma ΔC^*_{ab} 3.04, and in total colour difference ΔE^*_{ab} 6.28 (see Figure A.4 for a visual representation). Overall, the centre of the screen was found to have a slightly better uniformity than the surround. The ΔE^* was more affected by lightness non-uniformity than chroma non-uniformity.

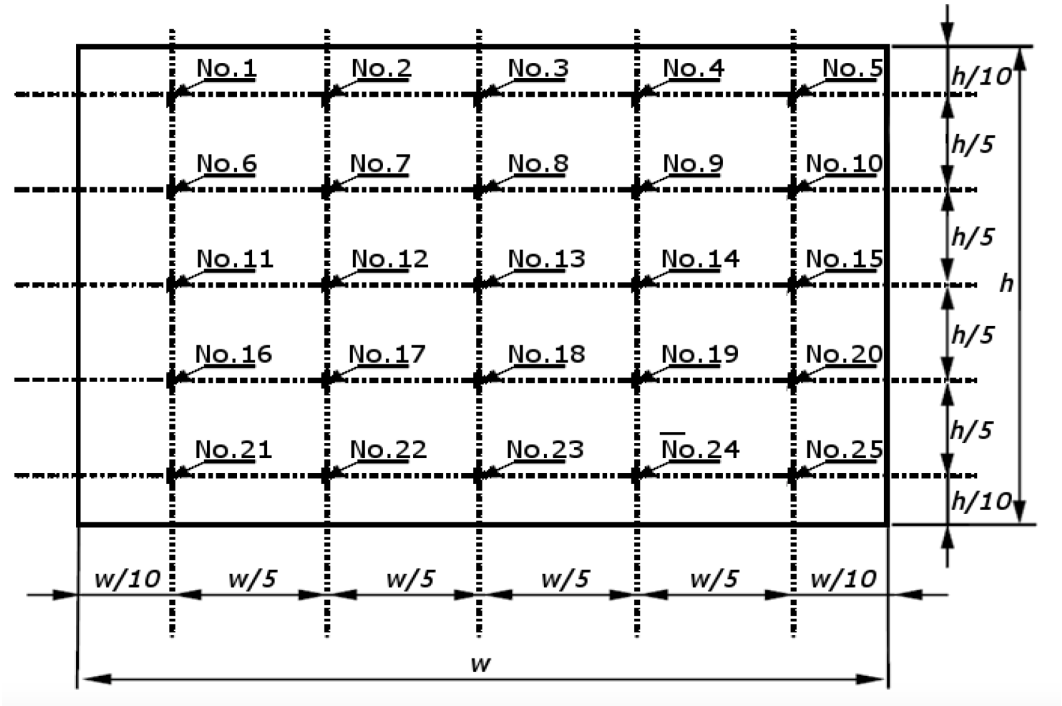


Figure A.3: The 25 measurement positions (or points) for monitor uniformity. Where h and w stand for monitor height and width respectively. Adopted from BS EN 61966-4 standard [299].

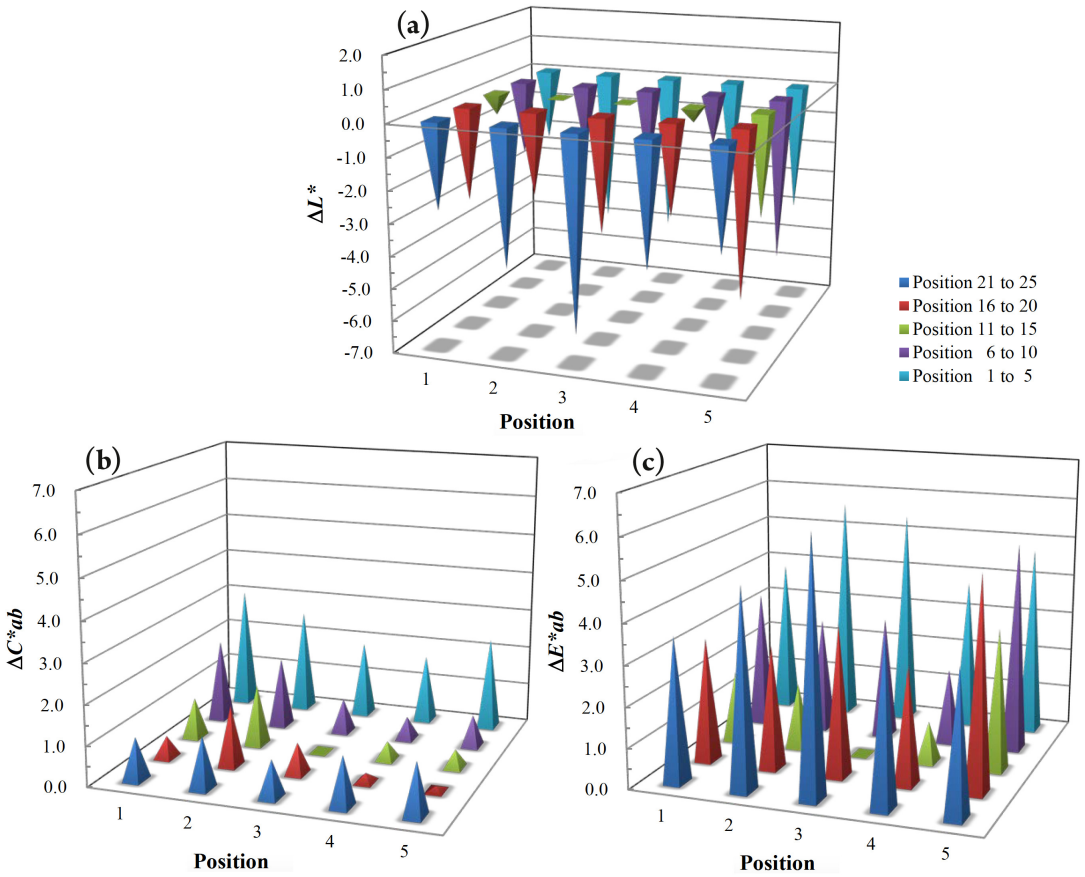


Figure A.4: LCD spatial uniformity measurements. Graphs a, b and c present differences in lightness (ΔL^*), chroma ΔC^*_{ab} and total colour ΔE^*_{ab} respectively from the reference position to the 25 positions across the display. Adopted from Park (2014) [298].

The measured spatial uniformity of the EIZO ColorEdge CG210 21.3" LCD was rather poor. Overall the second investigation with the samples from the CCTV DVRs would not have been affected greatly because the samples were very small and were placed in the middle of the LCD (see Figure A.6). In the first investigation, the visual representation of the samples may have been affected, as the scenes were displayed across the entire LCD faceplate (see Figure A.5). However, in most scenes the areas of interest (the faces displayed in grey squares) were not placed at the very edges of the LCD.

Overall the spatial non-uniformity, although poor in terms of accurate colorimetry, would affect more image fidelity and image quality experiments than image usefulness tests. How useful an image is for a recognition task depends on the information that image conveys (with respect to the original

in the specific experiment), not the accuracy of the information. A ΔE^* of 1 is the ultimate threshold of visibility for solid patches, where as a ΔE^* of 3 to 6 is still considered acceptable for the display of complex images [302–304].

Figure A.5: Example of the first psychophysical experiment. The figure is showing what was displayed on the entire LCD screen.

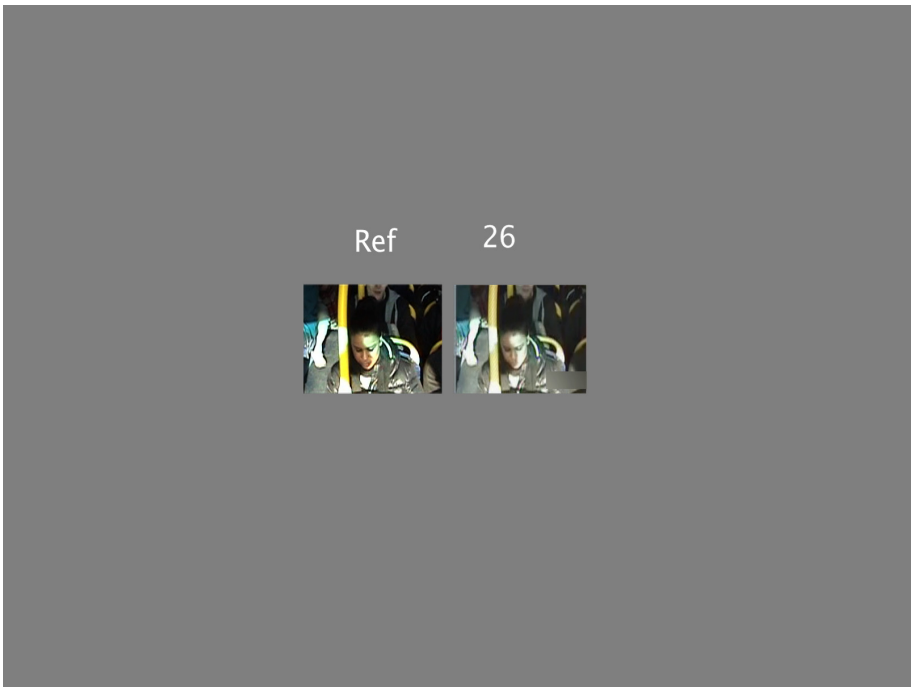


Figure A.6: Example of the second psychophysical experiment. The figure is showing what was displayed on the entire LCD screen.

- *Viewing angle characteristics*

  This measurement examines the effect on luminance Y and u', v' chromaticity coordinates of different display viewing angles. The LCD device was mounted with a tilted stand that allowed adjustable changes in vertical viewing angles. A rotating disk was placed under the LCD device to allow adjustable changes in horizontal viewing angles. The test samples consisted of 8 neutral and 3 pure primary colour patches. They were all displayed individually on the screen on a black background (according to the BS EN 61966- 4 standard [299]). Luminance (Y) and chromaticity coordinates (CIE 1976 u', v') measurements were performed horizontally +/- 40° (from the centre of the screen) at 10° interval and vertically between 0° to +20° and at 5° interval. The results are plotted in Figures A.7 to A.9.

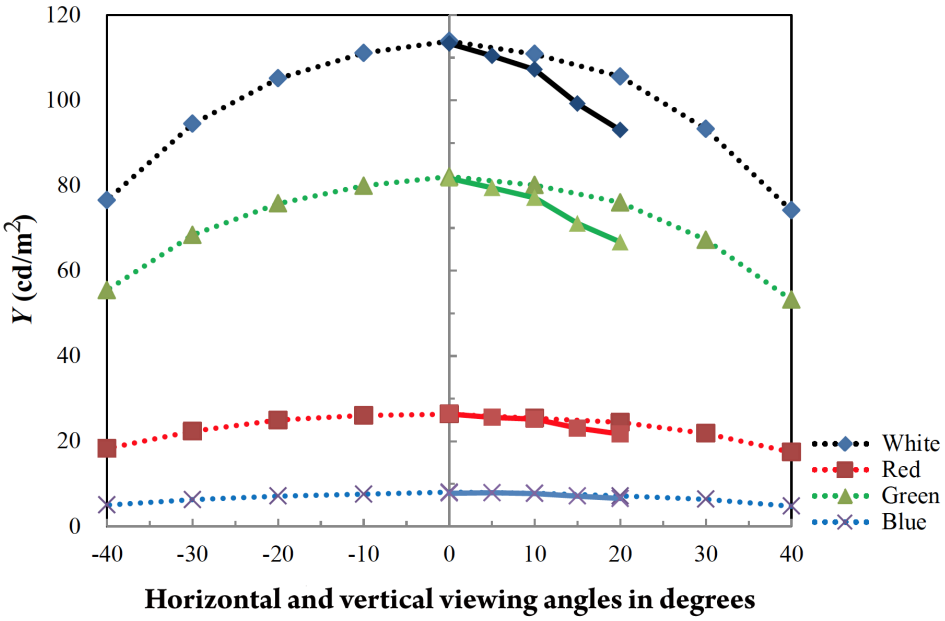


Figure A.7: The effect of different viewing angles to pure primaries (red, green and blue) and the white. Solid lines present results from the vertical luminance and broken lines of horizontal luminance. Adopted from Park (2014) [298].

The luminance value has decreased for the white and pure red patches and that decrease was slightly higher for the vertical angles (see Figures A.7). A similar behaviour can be observed from the results obtained from the neutral patches (see Figure A.9). The changes of the chromaticity coordinate measurements

from the different viewing angles were small (see Figure A.8). The observers in the investigation in Chapter 4 were allowed to get closer to the screen or further away and change viewing angles. This is a common practice with police officers. The angular variations reported in this section would not thus affect the reported usefulness, as the officers would have the chance to check and double check the images from all viewing angles.

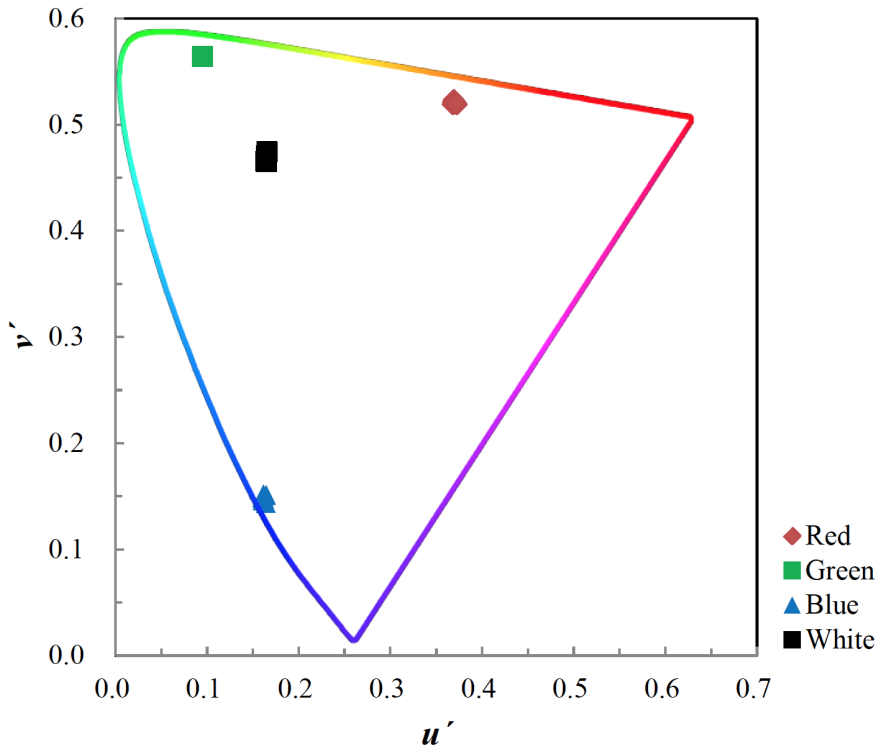


Figure A.8: The effect of different viewing angles to chromaticity measurements. Adopted from Park (2014) [298].

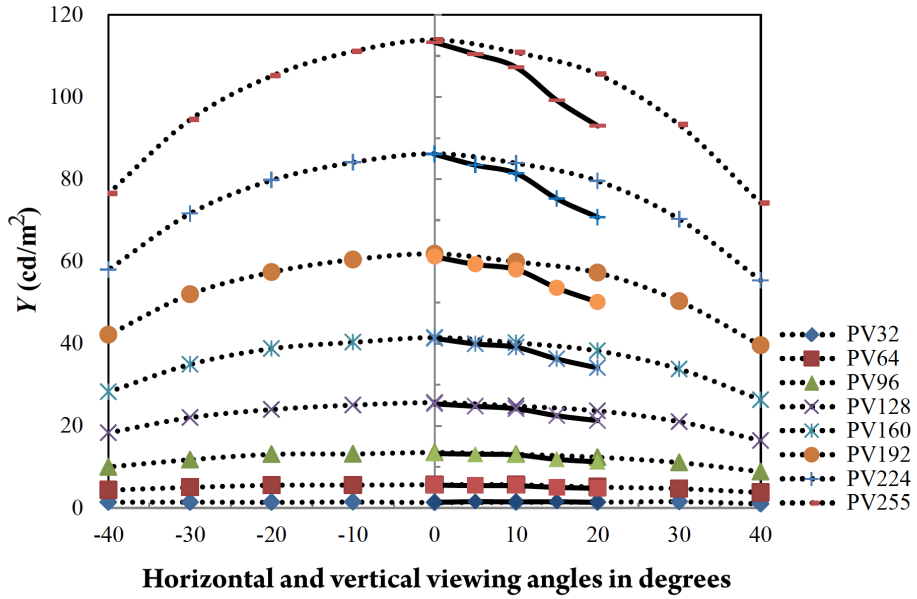


Figure A.9: The effect of different viewing angles to neutral patches. Solid lines present results from the vertical luminance and broken lines of horizontal luminance. Adopted from Park (2014) [298].

The limitations of employing a non-contact measurement instrument (in this case the Konica-Minolta CS-200 tele-chroma meter) to measure angular variations are the reflections caused by the objects in the room and the light emitting from the measuring device. Provided that these were all controlled (i.e. measurement in total darkness, as reported) these should not have an effect. Non-contact measurements enable the independent capture of color information and this is applicable for any spatial point within its field of view [305].

APPENDIX B

Logistic Regression Analysis

Sys.A:Property	α	**std**	β	**std**	p
Distance					
$Close_{25fps}$	0.1601	1.5123	0.2456	0.2279	0.282
$Close_{5fps}$	-0.4694	1.3503	0.2818	0.2034	0.167
$Medium_{25fps}$	-169.30	18513.16	32.09	3418.17	0.993
$Medium_{5fps}$	-180.36	11814.70	33.58	2181.40	0.988
Far_{25fps}	-166.20	13060.88	31.24	2411.49	0.99
Far_{5fps}	-1.6919	2.1609	0.6671	0.3384	0.050.
Orientation					
$Diag_{25fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	0
$Diag_{5fps}$	-168.30	20009.78	31.59	3694.50	0.009
$Perp_{25fps}$	0.5262	1.3700	0.3397	0.2089	0.104
$Perp_{5fps}$	-1.4215	1.0407	0.5348	0.1604	0.000***
Description					
$OnePerson_{25fps}$	0.3087	1.4718	0.4043	0.2258	0.0739.
$OnePerson_{5fps}$	-1.5487	1.0800	0.5727	0.1671	0.000***
$TwoPeople_{25fps}$	2.398e+00	3.776e+00	-8.535e-16	5.604e-01	1.000
$TwoPeople_{5fps}$	0.06048	3.35164	0.32782	0.50931	0.522
Contrast					
$High_{25fps}$	-0.9834	2.7767	0.3212	0.4165	0.444
$High_{5fps}$	-0.6891	2.5405	0.1941	0.3778	0.610
$Medium_{25fps}$	-3.4502	3.7866	1.3064	0.6233	0.0366*
$Medium_{5fps}$	-2.0619	1.7447	0.7592	0.2735	0.005**
Low_{25fps}	0.3338	2.1703	0.2326	0.3257	0.0476
Low_{5fps}	-2.9603	1.9544	0.7039	0.3009	0.020*
Approach					
$Walk_{25fps}$	2.683279	2.644227	-0.006605	0.392249	0.987
$Walk_{5fps}$	0.2022	2.2934	0.3383	0.3492	0.334
Run_{25fps}	-0.1799	2.2881	0.3246	0.3470	0.351
Run_{5fps}	-3.9031	1.6617	0.7417	0.2535	0.004**
$CrouchRun_{25fps}$	0.8033	3.4253	0.2562	0.5183	0.623
$CrouchRun_{5fps}$	-2.2235	2.3982	0.5005	0.3634	0.173
$CreepWalk_{25fps}$	35.584	17.689	-4.205	2.362	0.078.
$CreepWalk_{5fps}$	2.4189	6.8267	0.4217	1.0531	0.690
$Crawl_{25fps}$	-170.63	22297.11	31.97	4116.82	0.994
$Crawl_{5fps}$	-172.27	24024.25	32.23	4435.71	0.994
$CrouchWalk_{25fps}$	-180.40	24216.57	33.59	4471.22	0.994
$CrouchWalk_{5fps}$	-176.8	24618.5	33.0	4545.4	0.994
$BodyDrag_{25fps}$	-172.23	26026.05	32.58	4805.31	0.995
$BodyDrag_{5fps}$	2.812e+01	1.380e+00	1.482e-07	2.049e-01	1
$LogRoll_{25fps}$	2.812e+01	2.183e+00	1.443e-08	3.240e-01	1
$LogRoll_{5fps}$	2.812e+01	2.183e+00	1.443e-08	3.240e-01	1
$WalkLadder_{25fps}$	-175.76	27640.13	32.83	5103.33	0.995
$WalkLadder_{5fps}$	-173.86	27862.34	32.85	5144.35	0.995

Table B.1: Information on the fitted logistic models in Figures 6.6 and 6.7 for detailed performance of System A (values obtained as in Table 6.3 for each individual scene property).

Sys.B:Property	α	**std**	β	**std**	p
Distance					
$Close_{25fps}$	1.00368	1.37167	0.08466	0.20455	0.679
$Close_{5fps}$	0.26766	1.19663	0.07929	0.17805	0.657
$Medium_{25fps}$	2.4113	1.7800	-0.0463	0.2634	0.861
$Medium_{5fps}$	1.64653	1.45063	0.02234	0.21559	0.918
Far_{25fps}	2.2175	1.6005	-0.0546	0.2367	0.818
Far_{5fps}	-0.1935	1.4074	0.2785	0.2124	0.191
Orientation					
$Diag_{25fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
$Diag_{5fps}$	2.812e+01	1.044e+00	-4.925e-0	1.549e-01	1
$Perp_{25fps}$	1.645760	0.900750	0.005832	0.133731	0.9652
$Perp_{5fps}$	0.4055	0.7719	0.1190	0.1152	0.302
Description					
$OnePerson_{25fps}$	1.641278	0.901557	0.006694	0.133858	0.9601
$OnePerson_{5fps}$	0.3965	0.7724	0.1205	0.1153	0.296
$TwoPeople_{25fps}$	12.3287	11.2502	-0.8324	1.5814	0.600
$TwoPeople_{5fps}$	622.64	71135.87	-81.29	9358.87	0.993
Contrast					
$High_{25fps}$	6.6412	3.6331	-0.6942	0.5190	0.1868
$High_{5fps}$	1.08461	2.87765	0.03153	0.42662	0.941
$Medium_{25fps}$	1.4862	1.1296	0.0556	0.1678	0.741
$Medium_{5fps}$	0.5866	0.9415	0.1198	0.1402	0.393
Low_{25fps}	1.19580	1.95089	0.07127	0.28993	0.806
Low_{5fps}	0.2820	1.7983	0.1423	0.2679	0.596
Approach					
$Walk_{25fps}$	13.1728	11.2037	-0.8318	1.5748	0.598
$Walk_{5fps}$	628.45	77829.51	-81.94	10239.51	0.994
Run_{25fps}	3.0617	2.3347	-0.1244	0.3436	0.718
Run_{5fps}	-1.3913	1.5460	0.3209	0.2316	0.168
$CrouchRun_{25fps}$	-9.2296	3.4873	1.9689	0.6008	0.001**
$CrouchRun_{5fps}$	-0.4627	2.2929	0.2019	0.3427	0.558
$CreepWalk_{25fps}$	2.812e+01	9.307e-01	-4.237e-08	1.381e-01	1
$CreepWalk_{5fps}$	2.812e+01	9.307e-01	-4.237e-08	1.381e-01	1
$Crawl_{25fps}$	0.1998	1.8800	-0.1156	0.2798	0.681
$Crawl_{5fps}$	-0.7580	1.8672	0.0107	0.2770	0.969
$CrouchWalk_{25fps}$	2.4898	5.8243	0.4742	0.9033	0.602
$CrouchWalk_{5fps}$	12.0430	11.2786	-0.8327	1.5854	0.602
$BodyDrag_{25fps}$	0.2432	2.2391	-0.1338	0.3335	0.690
$BodyDrag_{5fps}$	-0.36872	1.77646	-0.03742	0.26391	0.888
$LogRoll_{25fps}$	6.931e-01	4.630e+00	2.002e-15	6.872e-01	1.000
$LogRoll_{5fps}$	-1.1409	4.3734	0.2565	0.6528	0.700
$WalkLadder_{25fps}$	8.7606	5.3000	-0.7407	0.7495	0.330
$WalkLadder_{5fps}$	-176.10	29547.94	33.06	5455.57	0.995

Table B.2: Information on the fitted logistic models in Figures 6.8 and 6.9 for detailed performance of System B (values obtained as in Table 6.3 for each individual scene property).

Sys.C:Property	α	**std**	β	**std**	p
Distance					
$Close_{25fps}$	3.2354	1.5227	-0.2477	0.2230	0.2679
$Close_{5fps}$	3.9245	1.7093	-0.3204	0.2491	0.1996
$Medium_{25fps}$	2.61729	1.87703	-0.07694	0.27717	0.782
$Medium_{5fps}$	2.283901	1.919788	0.004279	0.285012	0.988
Far_{25fps}	2.6038	1.6772	-0.1438	0.2469	0.561
Far_{5fps}	2.0782	1.6602	-0.0589	0.2456	0.811
Orientation					
$Diag_{25fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
$Diag_{5fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
$Perp_{25fps}$	2.7491	0.9703	-0.1691	0.1427	0.236
$Perp_{5fps}$	2.7116	1.0134	-0.1417	0.1491	0.342
Description					
$OnePerson_{25fps}$	2.7491	0.9703	-0.1691	0.1427	0.236
$OnePerson_{5fps}$	2.7116	1.0134	-0.1417	0.1491	0.342
$TwoPeople_{25fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
$TwoPeople_{5fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
Contrast					
$High_{25fps}$	2.079e+00	4.039e+00	-3.626e-15	5.980e-01	1.000
$High_{5fps}$	2.079e+00	4.039e+00	-3.626e-15	5.980e-01	1.000
$Medium_{25fps}$	3.4684	1.2494	-0.2364	0.1825	0.196
$Medium_{5fps}$	2.9318	1.3228	-0.1337	0.1943	0.4918
Low_{25fps}	1.75305	1.91797	-0.06574	0.28311	0.817
Low_{5fps}	2.8315	1.9453	-0.2039	0.2852	0.476
Approach					
$Walk_{25fps}$	2.812e+01	6.777e-01	-6.992e-08	1.006e-01	1
$Walk_{5fps}$	2.812e+01	6.777e-01	-6.992e-08	1.006e-01	1
Run_{25fps}	0.8336	1.5629	-0.1335	0.2320	0.566
Run_{5fps}	0.40221	1.56284	-0.05959	0.23194	0.798
$CrouchRun_{25fps}$	1.5722	2.0806	-0.2111	0.3083	0.496
$CrouchRun_{5fps}$	3.0178	2.2090	-0.3576	0.3247	0.275
$CreepWalk_{25fps}$	15.019	9.522	-1.596	1.307	0.225
$CreepWalk_{5fps}$	7.9698	5.5188	-0.5586	0.7875	0.480
$Crawl_{25fps}$	20.571	10.327	-2.431	1.400	0.087.
$Crawl_{5fps}$	35.173	19.749	-4.347	2.638	0.1044
$CrouchWalk_{25fps}$	2.812e+01	1.211e+00	-1.072e-08	1.797e-01	1
$CrouchWalk_{5fps}$	2.812e+01	1.211e+00	-1.072e-08	1.797e-01	1
$BodyDrag_{25fps}$	17.226	9.952	-1.922	1.358	0.1645
$BodyDrag_{5fps}$	640.70	91326.85	-83.74	12015.26	0.994
$LogRoll_{25fps}$	2.812e+01	2.183e+00	1.443e-08	3.240e-01	1
$LogRoll_{5fps}$	2.812e+01	2.183e+00	1.443e-08	3.240e-01	1
$WalkLadder_{25fps}$	2.812e+01	1.497e+00	9.209e-09	2.222e-01	1
$WalkLadder_{5fps}$	2.812e+01	1.497e+00	9.209e-09	2.222e-01	1

Table B.3: Information on the fitted logistic models in Figures 6.10 and 6.11 for detailed performance of System C (values obtained as in Table 6.3 for each individual scene property).

Sys.D:Property	α	**std**	β	**std**	p
Distance					
$Close_{25fps}$	2.2822	2.0312	0.2351	0.3076	0.446
$Close_{5fps}$	0.1789	1.3277	0.3255	0.2018	0.108
$Medium_{25fps}$	9.4012	7.4828	-0.4107	1.0766	0.703
$Medium_{5fps}$	-4.1278	2.6091	1.2916	0.4349	0.003**
Far_{25fps}	2.64255	2.55536	0.04369	0.38054	0.909
Far_{5fps}	-0.1636	2.2455	0.4600	0.3461	0.185
Orientation					
$Diag_{25fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
$Diag_{5fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
$Perp_{25fps}$	2.96314	1.80613	0.09493	0.27017	0.725
$Perp_{5fps}$	-0.3182	1.0816	0.4664	0.1667	0.005**
Description					
$OnePerson_{25fps}$	2.96314	1.80613	0.09493	0.27017	0.725
$OnePerson_{5fps}$	-0.3182	1.0816	0.4664	0.1667	0.005**
$TwoPeople_{25fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
$TwoPeople_{5fps}$	2.812e+01	1.044e+00	-4.925e-08	1.549e-01	1
Contrast					
$High_{25fps}$	2.423e+00	3.932e+00	-5.251e-05	5.821e-01	1.00
$High_{5fps}$	-3.9580	3.4331	0.9836	0.5444	0.076.
$Medium_{25fps}$	3.930784	2.677967	0.001852	0.396511	0.996
$Medium_{5fps}$	0.1961	1.3846	0.4090	0.2113	0.053.
Low_{25fps}	0.05636	3.19782	0.60733	0.49751	0.224
Low_{5fps}	0.2286	2.6106	0.4156	0.3987	0.299
Approach					
$Walk_{25fps}$	2.812e+01	6.777e-01	-6.992e-08	1.006e-01	1
$Walk_{5fps}$	-162.82	15635.22	30.67	2886.80	0.992
Run_{25fps}	0.8422	2.1505	0.2614	0.3256	0.424
Run_{5fps}	-2.2617	1.4379	0.6194	0.2221	0.006**
$CrouchRun_{25fps}$	2.8011	3.2885	-0.1044	0.4847	0.830
$CrouchRun_{5fps}$	-1.2945	1.8145	0.3494	0.2731	0.205
$CreepWalk_{25fps}$	2.812e+01	9.307e-01	-4.237e-08	1.381e-01	1
$CreepWalk_{5fps}$	2.812e+01	9.307e-01	-4.237e-08	1.381e-01	1
$Crawl_{25fps}$	2.812e+01	1.091e+00	-2.317e-08	1.620e-01	1
$Crawl_{5fps}$	2.812e+01	1.091e+00	-2.317e-08	1.620e-01	1
$CrouchWalk_{25fps}$	2.812e+01	1.211e+00	-1.072e-08	1.797e-01	1
$CrouchWalk_{5fps}$	2.812e+01	1.211e+00	-1.072e-08	1.797e-01	1
$BodyDrag_{25fps}$	2.812e+01	1.380e+00	1.482e-07	2.049e-01	1
$BodyDrag_{5fps}$	2.812e+01	1.380e+00	1.482e-07	2.049e-01	1
$LogRoll_{25fps}$	2.812e+01	2.183e+00	1.443e-08	3.240e-01	1
$LogRoll_{5fps}$	-181.20	36520.94	34.08	6743.03	0.996
$WalkLadder_{25fps}$	2.812e+01	1.497e+00	9.209e-09	2.222e-01	1
$WalkLadder_{5fps}$	2.812e+01	1.497e+00	9.209e-09	2.222e-01	1

Table B.4: Information on the fitted logistic models in Figures 6.12 and 6.13 for detailed performance of System D (values obtained as in Table 6.3 for each individual scene property).

APPENDIX C

Linear Regression Analysis

System	α	**std**	β	**std**	p
$Sys.A_{25fps}$	9.174e-01	2.756e-02	1.119e-05	1.907e-05	0.583
$Sys.A_{5fps}$	8.042e-01	5.824e-02	6.456e-05	4.795e-05	0.249
$Sys.B_{25fps}$	7.822e-02	2.149e-03	1.209e-08	1.769e-06	0.995
$Sys.B_{5fps}$	7.043e-02	2.532e-03	1.669e-06	2.085e-06	0.424
$Sys.C_{25fps}$	7.955e-02	2.343e-03	-2.036e-06	1.929e-06	0.292
$Sys.C_{5fps}$	8.065e-02	2.221e-03	-1.585e-06	1.829e-06	0.386
$Sys.D_{25fps}$	8.852e-02	8.580e-04	2.181e-07	7.064e-07	0.758
$Sys.D_{5fps}$	8.367e-02	1.188e-03	2.260e-06	9.785e-07	0.0212*

Table C.1: Details of the fitted linear regression models for the overall performance in Figure C.1. The first column provides the system name and the type of the raw data (25fps or 5fps). The second and fourth columns provide information on the derived coefficients of each model (intercept and slope). Next columns provide the calculated standard error on the coefficients (std). Where p is the statistical value identifying any significant trends.

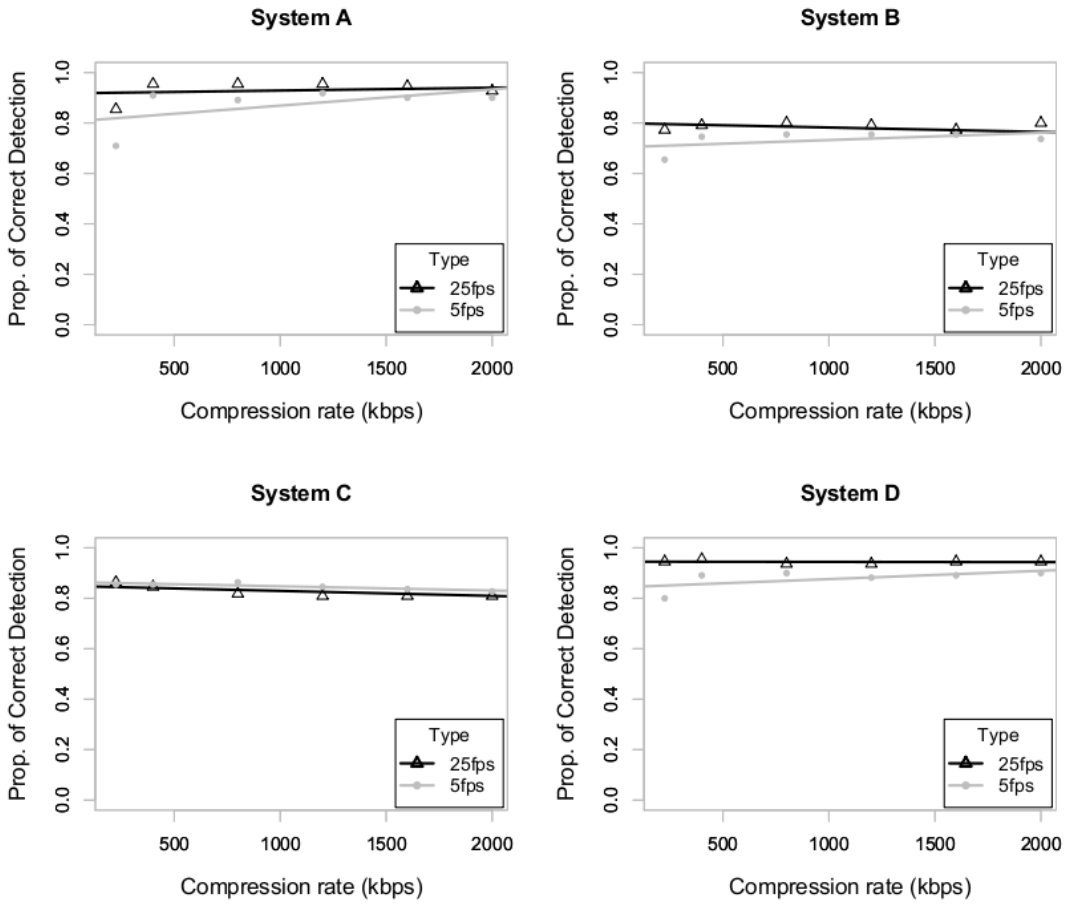


Figure C.1: Linear regression for the overall performance with respect to compression (in kbps) for systems A, B, C and D. Black triangles and black lines represent derived results from 25fps, and grey dots and grey lines represent derived results from 5fps. The lines are the obtained linear regression models from all the scenes and the points represent the always correctly identified scenes (the *Yes* scenes), both plotted against compression rate in kbps.

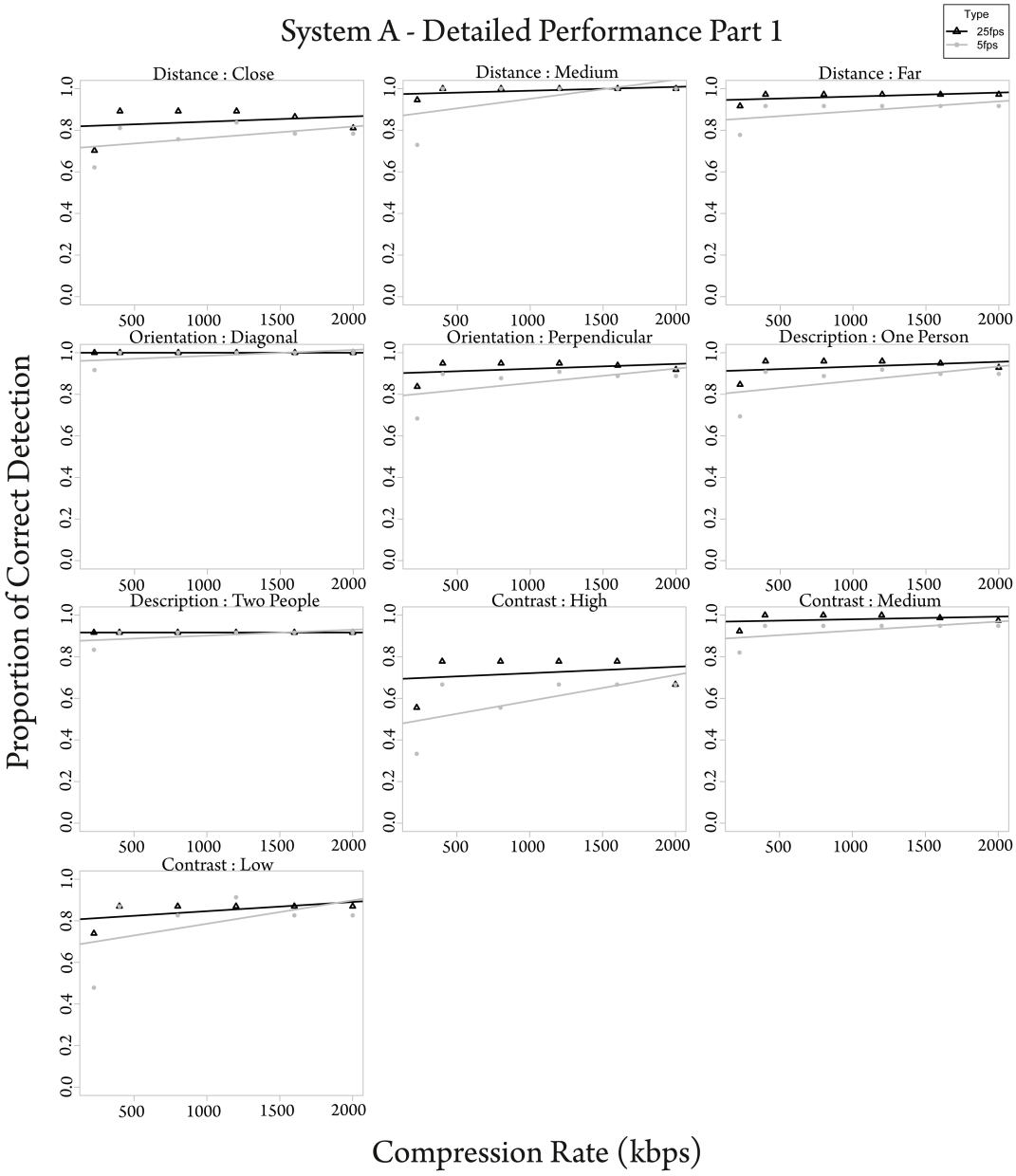


Figure C.2: Detailed performance with respect to compression (in kbps) for system A Part 1 (as graphs in Figure C.1).

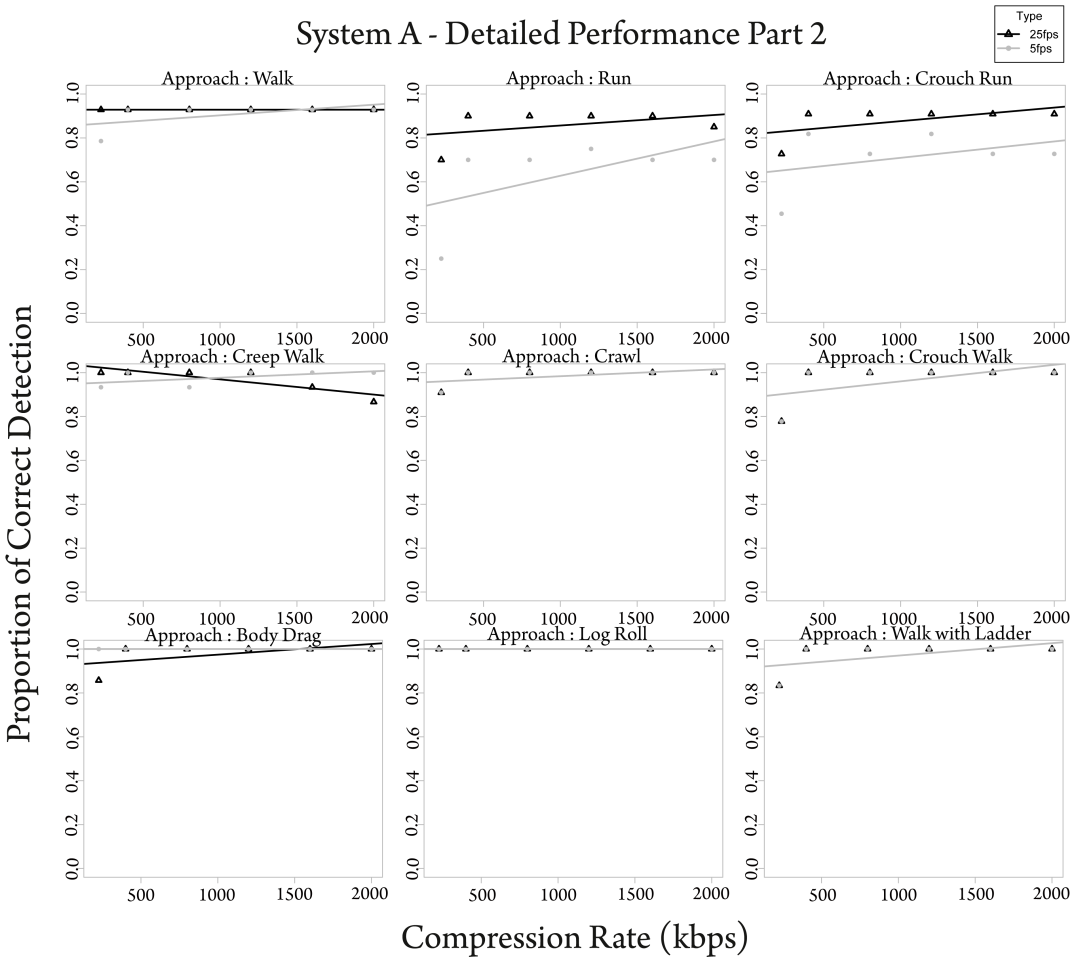


Figure C.3: Detailed performance with respect to compression (in kbps) for system A Part 2 (as graphs in Figure C.1).

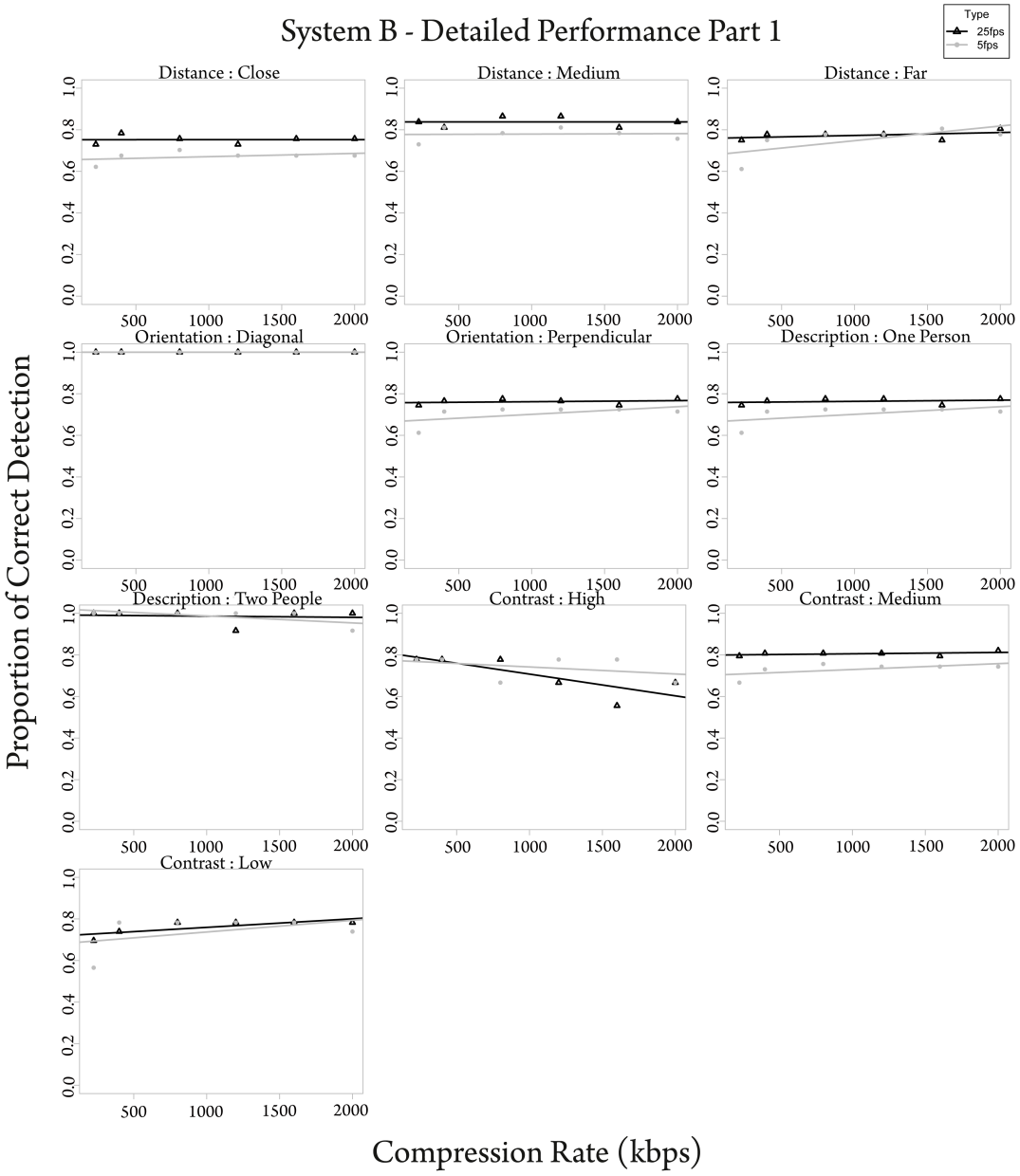


Figure C.4: Detailed performance with respect to compression (in kbps) for system B Part 1 (as graphs in Figure C.1).

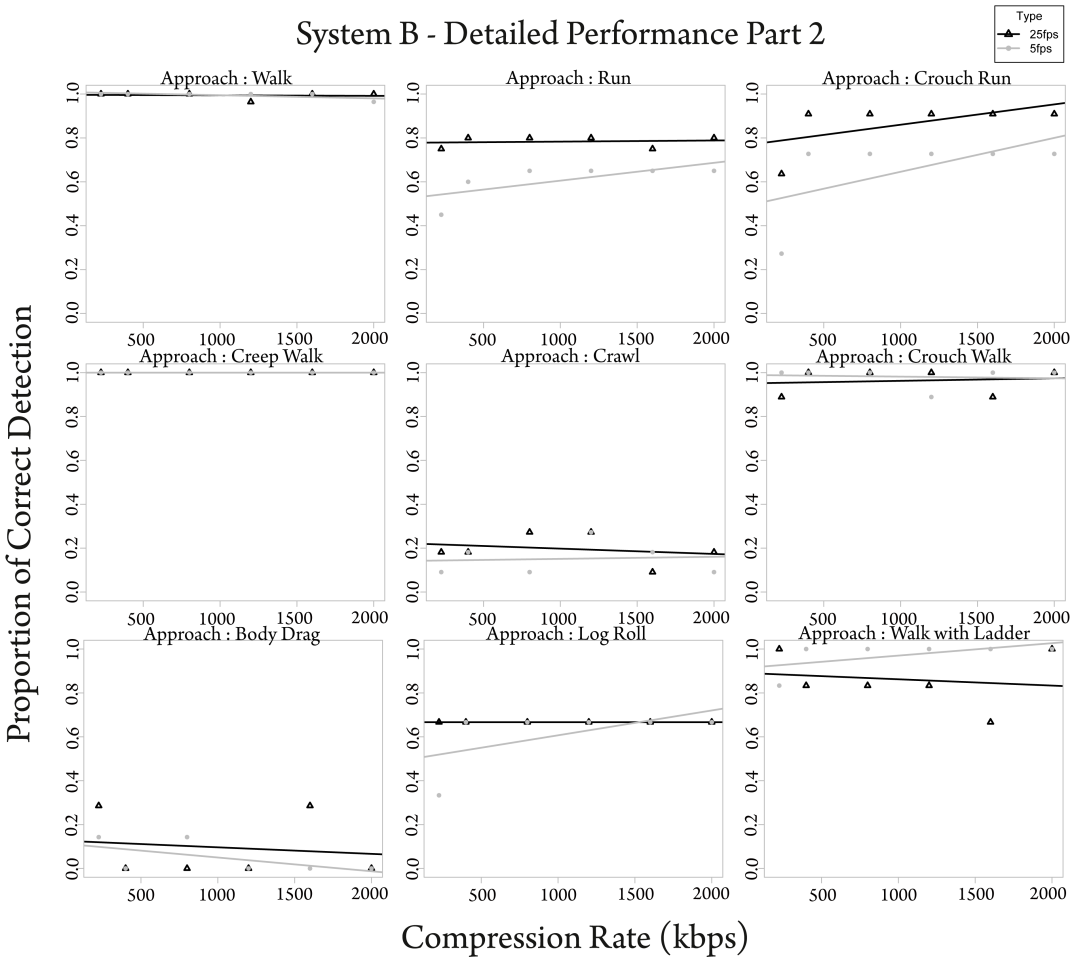


Figure C.5: Detailed performance with respect to compression (in kbps) for system B Part 2 (as graphs in Figure C.1).

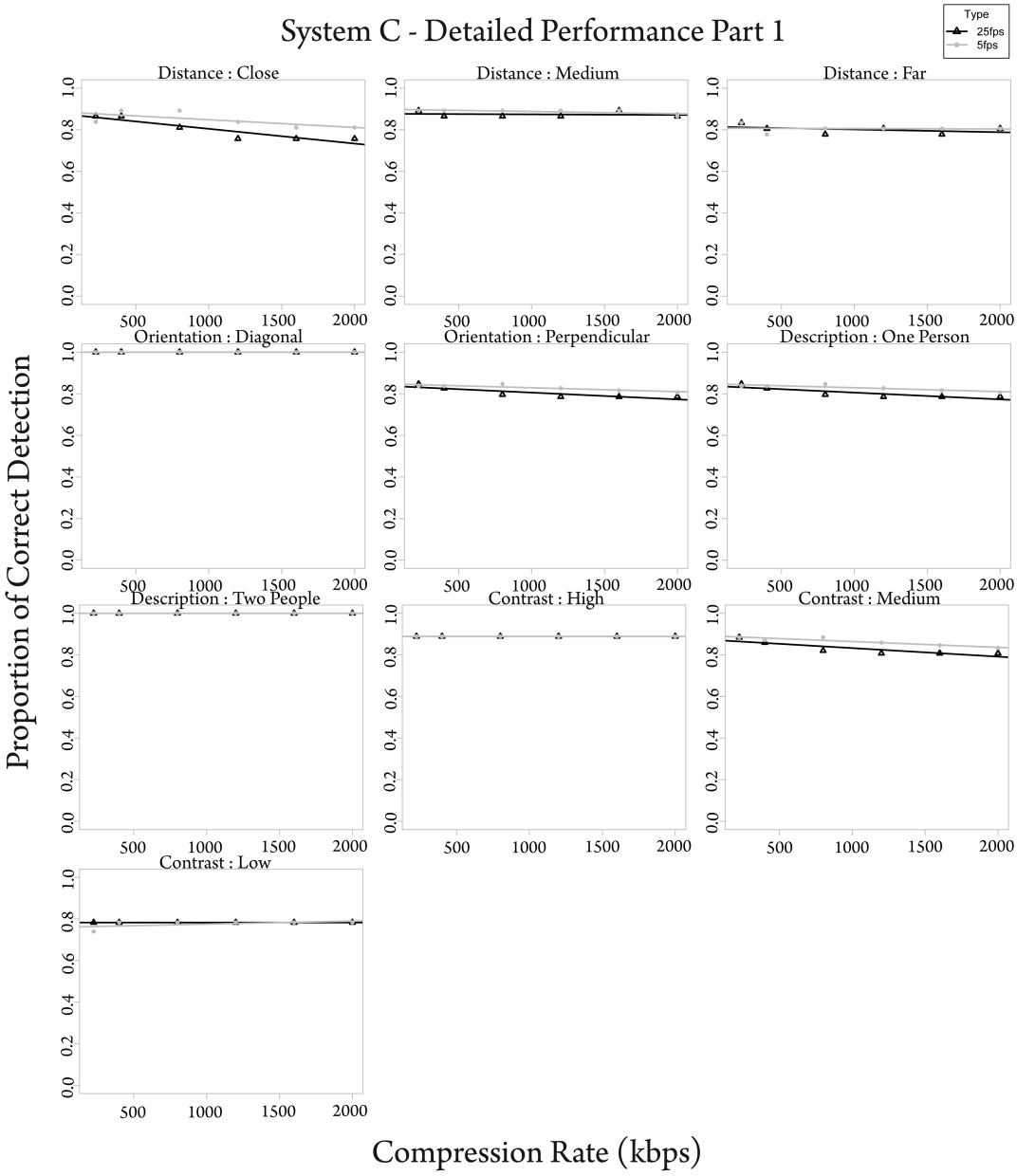


Figure C.6: Detailed performance with respect to compression (in kbps) for system C Part 1 (as graphs in Figure C.1).

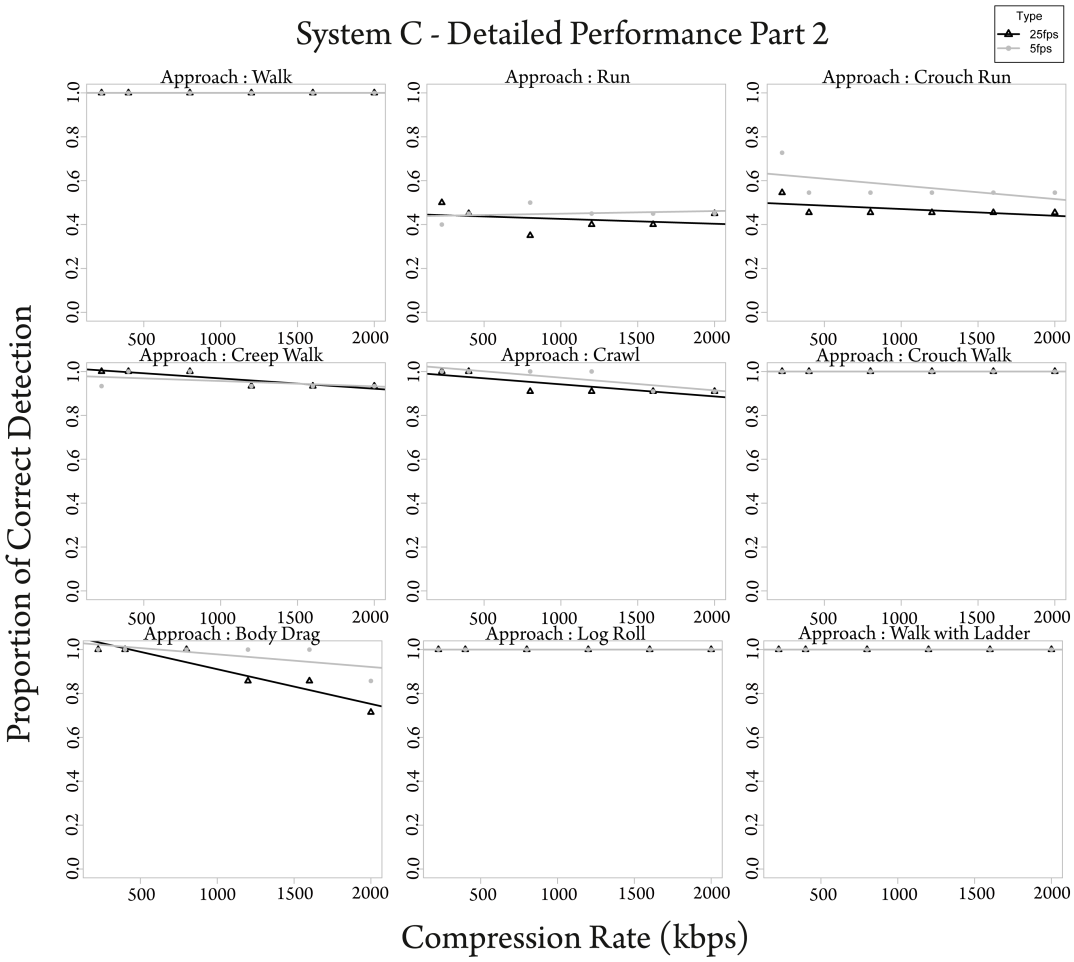


Figure C.7: Detailed performance with respect to compression (in kbps) for system C Part 2 (as graphs in Figure C.1).

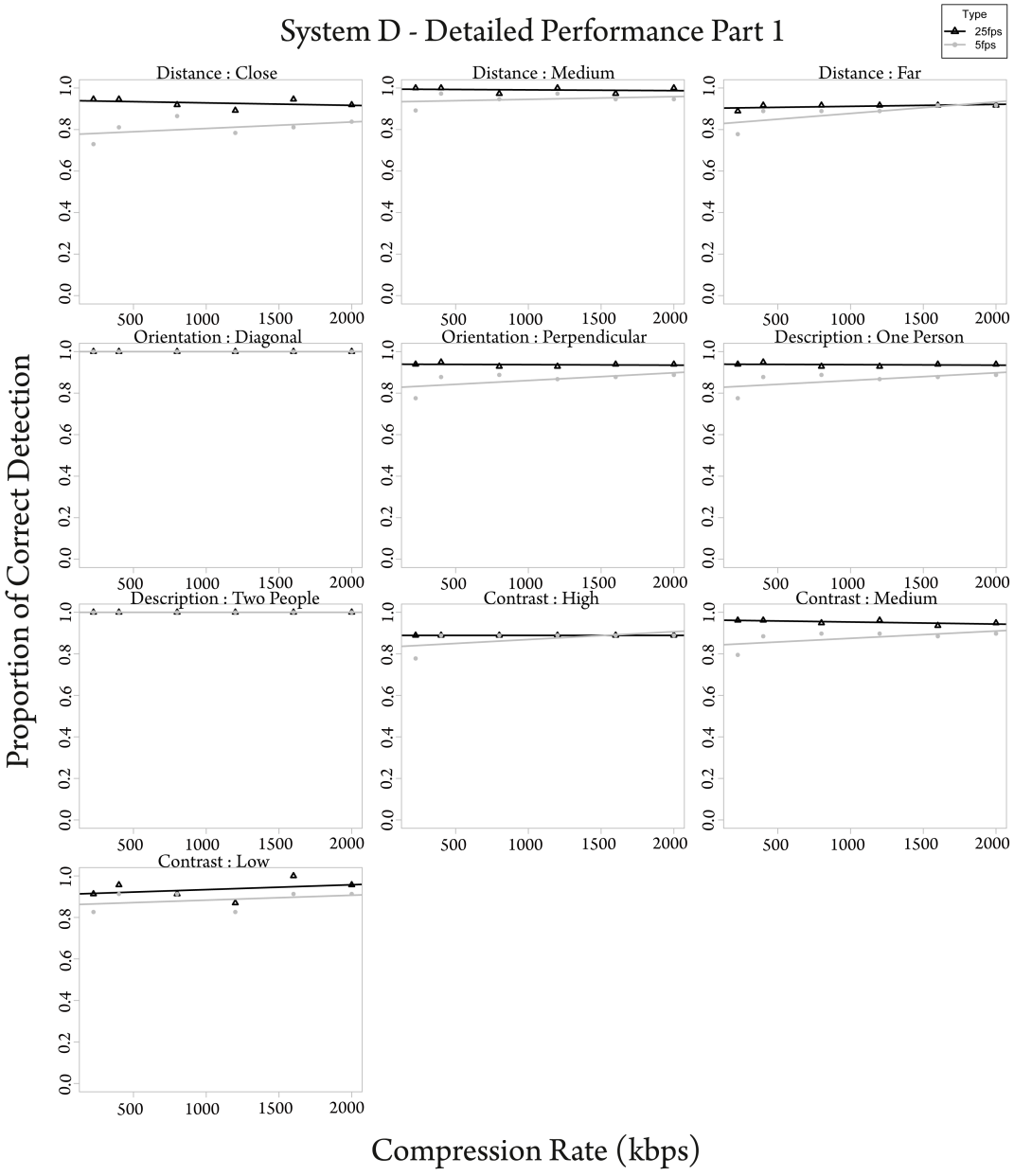


Figure C.8: Detailed performance with respect to compression (in kbps) for system D Part 1 (as graphs in Figure C.1).

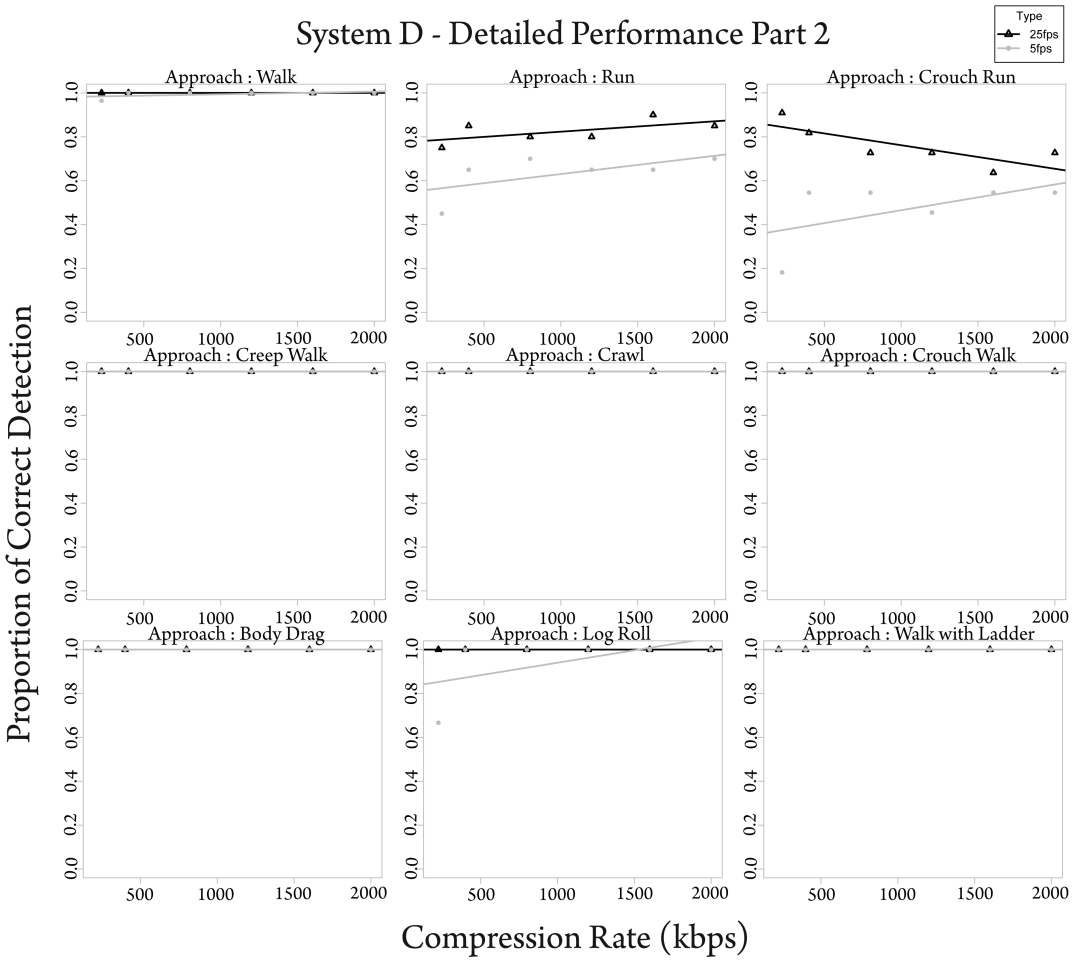


Figure C.9: Detailed performance with respect to compression (in kbps) for system D Part 2 (as graphs in Figure C.1).

Sys.A:Property	α	**std**	β	**std**	p
Distance					
$Close_{25fps}$	8.161e-01	6.432e-02	2.528e-05	5.296e-05	0.658
$Close_{5fps}$	7.095e-01	5.783e-02	5.423e-05	4.762e-05	0.318
$Medium_{25fps}$	9.719e-01	1.584e-02	1.836e-05	1.305e-05	0.232
$Medium_{5fps}$	8.597e-01	7.922e-02	9.180e-05	6.523e-05	0.232
Far_{25fps}	9.434e-01	1.628e-02	1.887e-05	1.341e-05	0.232
Far_{5fps}	8.446e-01	4.071e-02	4.717e-05	3.352e-05	0.232
Orientation					
$Diag_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Diag_{5fps}$	9.567e-01	2.443e-02	2.830e-05	2.011e-05	0.232
$Perp_{25fps}$	8.992e-01	3.604e-02	2.341e-05	2.967e-05	0.474
$Perp_{5fps}$	7.856e-01	6.241e-02	6.900e-05	5.138e-05	0.251
Description					
$OnePerson_{25fps}$	9.094e-01	3.604e-02	2.341e-05	2.967e-05	0.474
$OnePerson_{5fps}$	7.958e-01	6.241e-02	6.900e-05	5.138e-05	0.250
$TwoPeople_{25fps}$	0.9167	0.0000	0.0000	0.0000	NA
$TwoPeople_{5fps}$	8.734e-01	2.443e-02	2.830e-05	2.011e-05	0.232
Contrast					
$High_{25fps}$	6.903e-01	7.945e-02	3.077e-05	6.541e-05	0.663
$High_{5fps}$	4.637e-01	9.091e-02	1.243e-04	7.485e-05	0.172
$Medium_{25fps}$	9.675e-01	2.523e-02	1.280e-05	2.077e-05	0.571
$Medium_{5fps}$	8.822e-01	3.758e-02	4.354e-05	3.094e-05	0.232
Low_{25fps}	8.019e-01	3.823e-02	4.430e-05	3.148e-05	0.232
Low_{5fps}	6.732e-01	1.193e-01	1.125e-04	9.822e-05	0.316
Approach					
$Walk_{25fps}$	9.286e-01	2.139e-16	3.283e-19	1.761e-19	0.136
$Walk_{5fps}$	8.544e-01	4.188e-02	4.852e-05	3.448e-05	0.232
Run_{25fps}	8.087e-01	6.405e-02	4.781e-05	5.274e-05	0.416
Run_{5fps}	0.4712367	0.1359768	0.0001562	0.0001120	0.235
$CrouchRun_{25fps}$	8.147e-01	5.330e-02	6.175e-05	4.388e-05	0.232
$CrouchRun_{5fps}$	6.347e-01	1.084e-01	7.458e-05	8.925e-05	0.450
$CreepWalk_{25fps}$	1.039e+00	2.502e-02	-6.932e-05	2.060e-05	0.0282*
$CreepWalk_{5fps}$	9.474e-01	2.445e-02	2.926e-05	2.013e-05	0.220
$Crawl_{25fps}$	9.528e-01	2.665e-02	3.088e-05	2.194e-05	0.232
$Crawl_{5fps}$	9.528e-01	2.665e-02	3.088e-05	2.194e-05	0.232
$CrouchWalk_{25fps}$	8.847e-01	6.514e-02	7.548e-05	5.363e-05	0.232
$CrouchWalk_{5fps}$	8.847e-01	6.514e-02	7.548e-05	5.363e-05	0.232
$BodyDrag_{25fps}$	9.258e-01	4.188e-02	4.852e-05	3.448e-05	0.232
$BodyDrag_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$LogRoll_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$LogRoll_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$WalkLadder_{25fps}$	9.135e-01	4.885e-02	5.661e-05	4.022e-05	0.232
hline $WalkLadder_{5fps}$	9.135e-01	4.885e-02	5.661e-05	4.022e-05	0.232

Table C.2: Information on the fitted linear models in Figures C.2 and C.3 for detailed performance of System A (values obtained as in Table C.1 for each individual scene property).

Sys.B:Property	α	**std**	β	**std**	p
Distance					
$Close_{25fps}$	7.521e-01	1.786e-02	1.412e-07	1.471e-05	0.993
$Close_{5fps}$	6.549e-01	2.130e-02	1.568e-05	1.754e-05	0.422
$Medium_{25fps}$	8.378e-01	2.122e-02	2.983e-19	1.747e-05	1.00
$Medium_{5fps}$	7.771e-01	2.771e-02	2.118e-06	2.282e-05	0.93
Far_{25fps}	7.585e-01	1.625e-02	1.408e-05	1.338e-05	0.352
Far_{5fps}	6.768e-01	4.440e-02	7.054e-05	3.656e-05	0.126
Orientation					
$Diag_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Diag_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Perp_{25fps}$	7.565e-01	1.182e-0	5.225e-06	9.733e-06	0.620
$Perp_{5fps}$	6.643e-01	3.201e-02	3.674e-05	2.635e-05	0.236
Description					
$OnePerson_{25fps}$	7.575e-01	1.269e-02	5.919e-06	1.045e-05	0.601
$OnePerson_{5fps}$	6.643e-01	3.201e-02	3.674e-05	2.635e-05	0.236
$TwoPeople_{25fps}$	9.920e-01	2.967e-02	-5.661e-06	2.443e-05	0.828
$TwoPeople_{5fps}$	1.021e+00	2.185e-02	-3.353e-05	1.799e-05	0.136
Contrast					
$High_{25fps}$	8.121e-01	4.813e-02	-1.045e-04	3.963e-05	0.0577.
$High_{5fps}$	7.757e-01	4.604e-02	-3.367e-05	3.790e-05	0.425
$Medium_{25fps}$	7.988e-01	7.498e-03	6.498e-06	6.174e-06	0.352
$Medium_{5fps}$	7.014e-01	2.272e-02	2.827e-05	1.870e-05	0.205
Low_{25fps}	7.182e-01	1.991e-02			
Low_{5fps}	6.807e-01	6.825e-02	5.634e-05	5.619e-05	0.373
Approach					
$Walk_{25fps}$	9.966e-01	1.272e-02	-2.426e-06	1.047e-05	0.828
$Walk_{5fps}$	1.009e+00	9.365e-03	-1.437e-05	7.711e-06	0.136
Run_{25fps}	7.779e-01	2.245e-02	5.225e-06	1.848e-05	0.791
Run_{5fps}	5.240e-01	5.011e-02	8.125e-05	4.126e-05	0.120
$CrouchRun_{25fps}$	7.779e-01	2.245e-02	5.225e-06	1.848e-05	0.791
$CrouchRun_{5fps}$	5.240e-01	5.011e-02	8.125e-05	4.126e-05	0.120
$CreepWalk_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$CreepWalk_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Crawl_{25fps}$	2.221e-01	5.825e-02	-2.423e-05	4.796e-05	0.640
$Crawl_{5fps}$	1.417e-01	6.491e-02	9.501e-06	5.345e-05	0.8675
$CrouchWalk_{25fps}$	9.509e-01	4.988e-02	1.161e-05	4.107e-05	0.791
$CrouchWalk_{5fps}$	9.893e-01	3.956e-02	-7.548e-06	3.257e-05	0.828
$BodyDrag_{25fps}$	1.262e-01	1.283e-01	-2.986e-05	1.056e-04	0.791
$BodyDrag_{5fps}$	1.127e-01	5.239e-02	-6.270e-05	4.314e-05	0.220
$LogRoll_{25fps}$	0.6667	0.0000	0.0000	0.0000	NA
$LogRoll_{5fps}$	0.6667	0.0000	0.0000	0.0000	NA
$WalkLadder_{25fps}$	8.909e-01	1.088e-01	-2.874e-05	8.955e-05	0.764
$WalkLadder_{5fps}$	9.135e-01	4.885e-02	5.661e-05	4.022e-05	0.232

Table C.3: Information on the fitted linear models in Figures C.4 and C.5 for detailed performance of System B (values obtained as in Table C.1 for each individual scene property).

Sys.C:Property	α	**std**	β	**std**	p
Distance					
$Close_{25fps}$	8.754e-01	1.799e-02	-7.090e-05	1.481e-05	0.009**
$Close_{5fps}$	8.852e-01	2.337e-02	-3.700e-05	1.924e-05	0.127
$Medium_{25fps}$	8.768e-01	1.213e-02	-2.825e-06	9.989e-06	0.791
$Medium_{5fps}$	8.987e-01	7.087e-03	-1.087e-05	5.835e-06	0.136
Far_{25fps}	8.146e-01	1.651e-02	-1.321e-05	1.360e-05	0.386
Far_{5fps}	8.077e-01	1.537e-02	-2.032e-06	1.266e-05	0.880
Orientation					
$Diag_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Diag_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Perp_{25fps}$	8.383e-01	1.153e-02	-3.269e-05	9.497e-06	0.026*
$Perp_{5fps}$	8.478e-01	6.576e-03	-1.882e-05	5.415e-06	0.025*
Description					
$OnePerson_{25fps}$	8.383e-01	1.153e-02	-3.269e-05	9.497e-06	0.026*
$OnePerson_{5fps}$	8.478e-01	6.576e-03	-1.882e-05	5.415e-06	0.025*
$TwoPeople_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$TwoPeople_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
Contrast					
$High_{25fps}$	8.889e-01	1.426e-16	2.188e-19	1.174e-19	0.136
$High_{5fps}$	8.889e-01	1.426e-16	2.188e-19	1.174e-19	0.136
$Medium_{25fps}$	8.738e-01	1.449e-02	-4.107e-05	1.193e-05	0.026*
$Medium_{5fps}$	8.923e-01	6.977e-03	-2.800e-05	5.744e-06	0.0082**
Low_{25fps}	7.826e-01	2.139e-16	3.283e-19	1.761e-19	0.136
Low_{5fps}	7.600e-01	1.274e-02	1.477e-05	1.049e-05	0.232
Approach					
$Walk_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Walk_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
Run_{25fps}	4.480e-01	4.402e-02	-2.221e-05	3.624e-05	0.573
Run_{5fps}	4.375e-01	2.679e-02	1.202e-05	2.206e-05	0.615
$CrouchRun_{25fps}$	5.017e-01	2.665e-02	-3.088e-05	2.194e-05	0.232
$CrouchRun_{5fps}$	6.398e-01	5.330e-02	-6.175e-05	4.388e-05	0.232
$CreepWalk_{25fps}$	1.015e+00	1.456e-02	-4.703e-05	1.199e-05	0.017*
$CreepWalk_{5fps}$	9.809e-01	2.635e-02	-2.438e-05	2.169e-05	0.324
$Crawl_{25fps}$	9.966e-01	2.406e-02	-5.510e-05	1.981e-05	0.050*
$Crawl_{5fps}$	1.030e+00	2.145e-02	-5.795e-05	1.766e-05	0.0305*
$CrouchWalk_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$CrouchWalk_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$BodyDrag_{25fps}$	1.069e+00	3.537e-02	-1.582e-04	2.913e-05	0.006**
$BodyDrag_{5fps}$	1.036e+00	3.746e-02	-5.748e-05	3.084e-05	0.136
$LogRoll_{25fps}$	1.000e+00	1.426e-1	2.188e-19	1.174e-19	0.136
$LogRoll_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$WalkLadder_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$WalkLadder_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136

Table C.4: Information on the fitted linear models in Figures C.6 and C.7 for detailed performance of System C (values obtained as in Table C.1 for each individual scene property).

Sys.D:Property hline **Distance**	α	**std**	β	**std**	p
$Close_{25fps}$	9.402e-01	1.798e-02	-1.186e-05	1.481e-05	0.468
$Close_{5fps}$	7.739e-01	3.621e-02	3.121e-05	2.981e-05	0.354
$Medium_{25fps}$	9.948e-01	1.205e-02	-3.672e-06	9.920e-06	0.730
$Medium_{5fps}$	9.325e-01	2.477e-02	1.299e-05	2.039e-05	0.559
Far_{25fps}	9.022e-01	8.142e-03	9.435e-06	6.704e-06	0.232
Far_{5fps}	8.221e-01	3.046e-02	5.545e-05	2.508e-05	0.092.
Orientation					
$Diag_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Diag_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Perp_{25fps}$	9.396e-01	6.585e-03	-2.399e-06	5.421e-06	0.681
$Perp_{5fps}$	8.238e-01	3.050e-02	3.706e-05	2.511e-05	0.214
Description					
$OnePerson_{25fps}$	9.396e-01	6.585e-03	-2.399e-06	5.421e-06	0.681
$OnePerson_{5fps}$	8.238e-01	3.050e-02	3.706e-05	2.511e-05	0.214
$TwoPeople_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$TwoPeople_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
Contrast					
$High_{25fps}$	8.889e-01	1.426e-16	2.188e-19	1.174e-19	0.136
$High_{5fps}$	8.312e-01	3.257e-02	3.774e-05	2.682e-05	0.232
$Medium_{25fps}$	9.633e-01	6.943e-03	-9.915e-06	5.717e-0	0.158
$Medium_{5fps}$	8.395e-01	2.814e-02	3.524e-05	2.317e-05	0.203
Low_{25fps}	9.105e-01	3.743e-02	2.340e-05	3.082e-05	0.490
Low_{5fps}	8.595e-01	3.672e-02	2.363e-05	3.023e-05	0.478
Approach					
$Walk_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Walk_{5fps}$	9.815e-01	1.047e-02	1.213e-05	8.620e-06	0.232
Run_{25fps}	7.759e-01	3.599e-02	4.729e-05	2.963e-05	0.186
Run_{5fps}	5.471e-01	6.430e-02	8.308e-05	5.294e-05	0.192
$CrouchRun_{25fps}$	8.690e-01	5.044e-02	-1.074e-04	4.153e-05	0.061.
$CrouchRun_{5fps}$	3.480e-01	1.062e-01	1.173e-04	8.742e-05	0.251
$CreepWalk_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$CreepWalk_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Crawl_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$Crawl_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$CrouchWalk_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$CrouchWalk_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$BodyDrag_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$BodyDrag_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$LogRoll_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$LogRoll_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$WalkLadder_{25fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136
$WalkLadder_{5fps}$	1.000e+00	1.426e-16	2.188e-19	1.174e-19	0.136

Table C.5: Information on the fitted linear models in Figures C.8 and C.9 for detailed performance of System D (values obtained as in Table C.1 for each individual scene property).

Abbreviations

AVC	Advanced Video Coding
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AFR	Automated Face Recognition
BDRT	Benton Face Recognition Test
bpp	bits per pixel
B-Frame	Biderectionally Predictive Frame
CRT	Cathode Ray Tube
CAST	Centre of Applied Science and Technology
CCD	Charge Coupled Device
CIE	Commission Internationale de l'Eclairage
CBR	Constant Bit Rate
DNA	Deoxyribonucleic Acid
DCR	Digital Camcorder
DV	Digital Video

DVR	Digital Video Recorder
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
ESF	Edge Spread Function
FR	Face Recognition
FRVT	Face Recognition Vendor Test
FFmpeg	Fast Forward moving picture expert group
fps	frames per second
FUN	Fidelity Usefulness Naturalness
FCP	Final Cut Pro
GMM	Gaussian Mixture Models
GDA	General Discriminant Analysis
GB	Gigabytes
HFR	Human Face Recognition
HVS	Human Visual System
ID	Identity Document
iCAM	Image Colour Appearance Model
iLIDS	Imagery Library for Intelligent Detection Systems
IEC	International Electrotechnical Commission
I-Frame	Inter Frame
ISO	International Organization for Standardisation
ITU	International Telecommunication Union
IP	Internet Protocol
JM	Joint Model
JPEG	Joint Photographic Experts Group
JND	Just Noticeable Difference
KFA	Kernel Fisher Analysis

kbps	kilobits per second
LDA	Linear Discriminant Analysis
LSF	Line Spread Function
LCD	Liquid Crystal Display
ML	Machine Learning
MSE	Mean Square Error
Mbits/s	Megabits per second
MPS	Metropolitan Police Service
MTF	Modulation Transfer Function
MC	Motion Compensation
MPEG	Moving Picture Experts Group
NIST	National Institute of Standards and Technology
NPIA	National Policing Improvement Agency
NTSC	National Television System Committee
NPR	Number Plate Recognition
OECF	Opto-Electronic Conversion Function
PSNR	Peak Signal to Noise Ratio
PAL	Phase Alternating Line
PSE	Point of Subjective Equality
PSF	Point Spread Function
PNG	Portable Network Graphics
PCA	Principal Component Analysis
P-Frame	Predictive Frame
QP	Quantisation Parameter
ROC	Receiver Operating Characteristic
RGB	Red Green Blue
RMSE	Root Mean Square Error

SECAM	Séquentiel Couleur Avec Mémoire
SNR	Signal to Noise Ratio
SFR	Spatial Frequency Response
std	standard deviation
SZ	Sterile Zone
TIFF	Tagged Image File Format
TRV	Target Recognition Video
TfL	Transport for London
3D	3-Dimensional
2D	2-Dimensional
UK	United Kingdom
VA	Video Analytics

Bibliography

[1] G. Hutton and D. Johnston, *Police manual evidence and procedure*, Blackstone Press Ltd (1998/99).

[2] M. Gill and A. Spriggs, "Assessing the impact of CCTV," Tech. Rep. Home Office Research Study 292, Home Office Research Development and Statistics Directorate (2005).

[3] "Royston Jackson v Regina, No 201002220 d4," tech. rep., Court of Appeal, Criminal Division (2011).

[4] M. S. Nixon, I. Bouchrika, B. Arbab-Zavar, and J. N. Carter, "On use of biometrics in forensics: gait and ear," in *European Signal Processing Conference*, (2010).

[5] P. Britton, "Video: Terrifying CCTV footage shows shopkeeper fighting back after robbers armed with machete storm Salford store," (2015).

[6] "Single Image Photogrammetry," tech. rep., Home Office Scientific Development Branch (2007).

[7] A. Criminisi, *Accurate Visual Metrology from Single and Multiple Uncalibrated Images*. PhD thesis, Robotics Research Group, Department of Engineering

Science, University of Oxford (1999).

[8] H. Komatsu, "The neural mechanisms of perceptual filling-in," *Nature Reviews Neuroscience* **7**, 220–231 (2006).

[9] A. M. Burton, S. Wilson, M. Cowan, and V. Bruce, "Face recognition in poor-quality video: Evidence from security surveillance," *Psychological Science* **10**, 243–248 (1999).

[10] V. Bruce, Z. Henderson, C. Newman, and A. Burton, "Matching identities of familiar and unfamiliar faces caught on CCTV images," *Journal of Experimental Psychology: Applied* **7**(3), 207–218 (2001).

[11] R. Kemp, N. Towell, and G. Pike, "When seeing should not be believing: Photographs, credit cards and fraud," *Applied Cognitive Psychology* **11**(3), 211–222 (1997).

[12] G. Davies and S. Thasen, "Closed-Circuit Television: How effective an identification aid?," *British Journal of Psychology* **91**(3), 411–426 (2000).

[13] A. Hillstrom, J. Sauer, and L. Hope, "Training methods for facial image comparison: a literature review." Working Paper. Stationery Office. (2011).

[14] "Subjective video quality assessment methods for recognition tasks," Rec. ITU-T P.912 in Series P: Terminals and subjective and objective assessment methods, ITU (2008).

[15] M. Klima, P. Pata, K. Fliegel, and P. Hanzlik, "Subjective image quality evaluation in security imaging systems," in *39th Annual 2005 International Carnahan Conference on Security Technology, 2005. CCST '05*, 19–22 (2005).

[16] "Information technlogy - Biometrics sample quality," Standard PD ISO/IEC TR 29749 - 5:2010, BSi (2010).

[17] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Quality measures in biometric systems," *IEEE Security Privacy* **10**(6), 52–62 (2012).

[18] S. N. Yendrikhovskij, "Image quality and colour characterisation," in *MacDonald, L. W., Luo, M. R., Colour image science - Exploring digital media*, 393–420, John Wiley and Sons, Ltd., Chichester, UK (2002).

[19] "Passport photo requirements: www.gov.uk/photos-for-passports," tech. rep., UK Government (2013).

[20] E. Bilissi, *Aspect of image quality and the internet.* PhD thesis, Image Technology Research Group, University of Westminster, Chapter 2 (2004).

[21] "Research project looking at CCTV imagery format structure, version 1," produced for the centre for applied science and technology, Sira Defence and Security Ltd (2011).

[22] "Onviftm streaming specification," streaming specification - version 2.10, ONVIF (2011).

[23] X. Jin and S. Goto, "Encoder adaptable difference detection for low power video compression in surveillance system," *Image Communication* **26**, 130–142 (2011).

[24] H. Kalva, "The H.264 video coding standard," *IEEE Multimedia* **13**(4), 86–90 (2006).

[25] N. Gagvani, "Challenges in video analytics," in *Embedded Computer Vision*, B. Kisačanin, S. Bhattacharyya, and S. Chai, Eds., *Advances in Pattern Recognition*, ch. 12, 237–256, Springer London (2009).

[26] C. Regazzoni, A. Cavallaro, Y. Wu, J. Konrad, and A. Hampapur, "Video analytics for surveillance: Theory and practice [from the guest editors]," *IEEE Signal Processing Magazine* **27**, 16–17 (2010).

[27] S. Runkin, "Private communications - iLIDS team in Home Office Science CAST." (2012).

[28] M. Klima, J. Pazderak, M. Bernas, P. Pata, J. Hozman, and K. Roubik, "Objective and subjective image quality evaluation for security technology,"

in *IEEE 35th International Carnahan Conference on Security Technology*, 108–114 (2001).

[29] M. Klima and V. Kloucek, "Some remarks on very high-rate image compression and its impact on security image data subjective evaluation," in *36th Annual 2002 International Carnahan Conference on Security Technology, 2002. Proceedings*, 198–201 (2002).

[30] M. Klima and K. Fliegel, "Image compression techniques in the field of security technology: examples and discussion," in *38th Annual 2004 International Carnahan Conference on Security Technology*, 278–284 (2004).

[31] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM* **34**, 30–44 (1991).

[32] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *Signal Processing Magazine, IEEE* **18**(5), 36–58 (2001).

[33] D. M. Blackburn, M. Bone, and P. J. Phillips, "Face recognition vendor test (FRVT) 2000, executive overview," Tech. Rep. A269514, National Institute of Justice (NIST) (2001).

[34] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms.," *Perception* **30**(3), 303–321 (2001).

[35] K. Delac, M. Grgic, and S. Grgic, "Effects of JPEG and JPEG2000 compression on face recognition," in *Pattern Recognition and Image Analysis*, S. Singh, M. Singh, C. Apte, and P. Perner, Eds., *Lecture Notes in Computer Science* **3687**, 136–145, Springer Berlin Heidelberg (2005).

[36] D. P. McGarry, C. M. Arndt, S. A. McCabe, and D. P. D'Amato, "Effects of compression and individual variability on face recognition performance," (2004).

[37] K. Delac, S. Grgic, and M. Grgic, "Image compression in face recognition - a literature survey," in *Recent advances in face recognition*, K. Delac, M. Grgic,

and M. S. Bartlett, Eds., ch. 1, 1 – 14, InTech (2008).

[38] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(10), 1090–1104 (2000).

[39] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing* **16**(5), 295 – 306 (1998).

[40] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(5), 831 – 846 (2010).

[41] P. Grother and M. Ngan, "Face recognition vendor test FRVT, performance of face identification algorithms," Report Interagency Report 8009, National Institute of Standards and Technology (NIST) (2014).

[42] G. Aggarwal, S. Biswas, P. Flynn, and K. Bowyer, "Predicting performance of face recognition systems: An image characterization approach," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 52 – 59 (2011).

[43] M. Sharkas and M. Elenien, "Eigenfaces vs. Fisherfaces vs. ICA for face recognition; a comparative study," in *9th International Conference on Signal Processing (ICSP)*, 914 – 919 (2008).

[44] A. Adler and T. Dembinsky, "Human vs. automatic measurement of biometric sample quality," in *Canadian Conference on Electrical and Computer Engineering, 2006. CCECE '06*, 2090–2093 (2006).

[45] C. Poppe, S. De Bruyne, P. Lambert, and R. Van de Walle, "Effect of H.264/AVC compression on object detection for video surveillance," in *10th Workshop on Image Analysis for Multimedia Interactive Services WIAMIS*, 129 – 132 (2009).

[46] P. Mahendrarajah, "Investigation of the performance of video analytics systems with compressed video using the i-lids sterile zone dataset," in *Proc. SPIE*, **8189**, 81890O–81890O–6 (2011).

[47] S. Triantaphillidou, E. Allen, and R. E. Jacobson, "Image Quality Comparison Between JPEG and JPEG2000. II. Scene Dependency, Scene Analysis, and Classification," *Journal of Imaging Science and Technology* **51** (2007).

[48] E. Allen, S. Triantaphillidou, and R. Jacobson, "Image quality comparison between JPEG and JPEG2000. i. psychophysical investigation," *Journal of Imaging Science and Technology* **51**(3), 248–258 (2007).

[49] "iLIDS dataset," Dataset, Home Office Centre for Applied Science and Technology, UK (2014).

[50] "Imagery library for intelligent detection systems, the iLIDS user guide," Guide 10/11, Home Office Centre for Applied Science and Technology, Crown Copyright (2011).

[51] V. Štruc, "The PHD face recognition toolbox," matlab toolbox, University of Ljubljana, Faculty of Electrotechnical Engineering, Slovenia (2012).

[52] H. Kruegle, "CCTV's Role in the Security Plan," in *CCTV surveillance video practices and technology*, 1–40, Butterworth-Heinemann (1995).

[53] M. Mccahill and C. Norris, *Estimating the Extent, Sophistication and Legality of CCTV in London.* Palgrave Macmillan, Basingstoke, Hampshire, England (2003).

[54] R. Thompson and G. Gerrard, "Two million cameras in the UK," *CCTV Image, official publication of the CCTV user group,* **42** (2011).

[55] M. McCahill and C. Norris, "Urbaneye: CCTV in London," Working Paper No.6, Centre for Criminology and Criminal Justice Centre for Criminology and Criminal Justice, University of Hull (2002).

[56] *Oxford Dictionary of English*, Oxford University Press, 3rd ed. (2010).

[57] *Cambridge Advanced Learner's Dictionary*, Cambridge University Press, 4th ed. (2013).

[58] P. J. Putter, "An investigation to ascertain whether Muzzle- prints of Cattle can be individualized by applying the same techniques as those used in Dactyliscopy," *Fingerprint Whorld* **6**(27), 55–59 (1982).

[59] L. Hesse, "The transition from video motion detection to intelligent scene discrimination and target tracking in automated video surveillance systems," *Security Journal* **15**, 69–78 (2002).

[60] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computer Surveys* **35**, 399–458 (2003).

[61] J. P. Davis and T. Valentine, "CCTV on trial: Matching video images with the defendant in the dock," *Applied Cognitive Psychology* **23**(4), 482–505 (2009).

[62] H. Hill, P. G. Schyns, and S. Akamatsu, "Information and viewpoint dependence in face recognition," *Cognition* **62**(2), 201 – 222 (1997).

[63] H. Hill and V. Bruce, "Effects of lighting on the perception of facial surfaces," *Journal of Experimental Psychology: Human Perception and Performance* **22**, 986–1004 (1996).

[64] A. J. O'Toole, S. Edelman, and H. H. Bulthoff, "Stimulus-specific effects in face recognition over changes in viewpoint.," *Vision Research* **38**, 2351–2363 (1998).

[65] V. S. Ramachandran, "Perceiving shape from shading," *Scientific American* **259**(2), 76–83 (1988).

[66] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: the problem of compensating for changes in illumination direction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **19**, 721–732 (1997).

[67] R. A. Smith, K. MacLennan-Brown, J. F. Tighe, N. Cohen, S. Triantaphillidou, and L. W. MacDonald, "Colour analysis and verification of CCTV images

under different lighting conditions," (2008).

[68] S. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**, 627–639 (2007).

[69] P. Vanezis, D. Lu, J. Cockburn, A. Gonzalez, G. McCombe, O. Trujillo, and M. Vanezis, "Morphological classification of facial features in adult caucasian males based on an assessment of photographs of 50 subjects.," *Journal of Forensic Sciences* **41**, 786–791 (1996).

[70] P. Vanezis and C. Brierley, "Facial image comparison of crime suspects using video superimposition," *Science and Justice* **36**, 27–33, 2015/01/24.

[71] R. Moreton and J. Morley, "Investigation into the use of photoanthropometry in facial image comparison," *Forensic Science International* **212**(1–3), 231 – 237 (2011).

[72] "Dean Atkins, Michael Atkins v The Queen, Case No. 200801604 d4," tech. rep., Court of Appeal, Criminal Division (1996).

[73] "R v Tang, Case No. NSWCCA 167," tech. rep., Court of Criminal Appeal NSW (2006).

[74] M. Bromby, "At face value? The use of facial mapping and CCTV image analysis for identification," *New Law Journal* **153**(7069), 302–304 (2003).

[75] "Best practice and face pose value documents, In: Best practice recommendation for the capture of mugshots, Version 2.0," tech. rep., NIST (1997).

[76] "Police standard for still digital image capture and data interchange of facial/mugshot and scar, mark and tattoo images," tech. rep., NPIA, London (2007).

[77] J. Ashok, V. Shivashankar, and P.V.G.S.Mudiraj, "An overview of biometrics," *International journal of computer science and engineering (IJCSE)* **2**(7), 2402–2408 (2010).

[78] R. M. Bolle, J. Connell, and S. Pankanti, "Introduction," in *Guide to biometrics*, ch. 1, 2 – 16, Springer (2003).

[79] S. Li and A. Jain, "Introduction," in *Handbook of Face Recognition*, 1–11, Springer New York (2005).

[80] B. Schouten and B. Jacobs, "Biometrics and their use in e-passports," *Image and Vision Computing* **27**(3), 305 – 312 (2009). Special Issue on Multimodal Biometrics Multimodal Biometrics Special Issue.

[81] W. Zhao and R. Chellappa, "A giuded tour of face processing," in *Face processing: advanced modeling and methods*, ch. 1, 3 – 53, Elsevier Inc (2006).

[82] R.-L. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 696 – 706 (2002).

[83] E. HjelmÃs and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding* **83**(3), 236 – 274 (2001).

[84] A. S. Tolba, A. El-Baz, and A. El-Harby, "Face recognition: A literature review," *International Journal of Information and Communication Engineering* (2:2), 88 – 103 (2006).

[85] R. Jafri and H. R. Arabnia, "A survey of face recognition techniques," *Journal of Information Processing Systems* **5**(2), 41 – 68 (2009).

[86] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: a review," *International Journal of Pattern Recognition and Artificial Intelligence* **26**(5) (2012).

[87] C. J. Mondloch, R. Le Grand, and D. Maurer, "Configural face processing develops more slowly than featural face processing.," *Perception* **31**(5), 553–566 (2002).

[88] Y. Konar, P. J. Bennett, and A. B. Sekuler, "Holistic processing is not correlated with face-identification accuracy," *Psychological Science* **21**, 38 – 43

(2010).

[89] G. Van Belle, P. De Graef, K. Verfaillie, T. Busigny, and B. Rossion, "Whole not hole: expert face recognition requires holistic perception.," *Neuropsychologia* **48**, 2620 – 2629 (2010).

[90] R. Kimchi and R. Amishav, "Faces as perceptual wholes: The interplay between component and configural properties in face processing," *Visual Cognition* **18**(7), 1034–1062 (2010).

[91] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice Hall Advanced Reference Series, Prentice Hall (1988).

[92] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Morgan Kaufmann, 2 ed. (1990).

[93] A. Billard, *Machine Learning Techniques*, Polycopiés de l'EPFL, EPFL (2010).

[94] D. Michie, D. Spiegelhalter, and C. Taylor, "Introduction," in *Machine Learning, Neural and Statistical Classification*, ch. 1, 1 – 5, MRC Biostatistics Unit, University Forvie Site, Cambridge (1994).

[95] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(12), 2930 – 2940 (2013).

[96] I. J. Cox, J. Ghosn, and P. N. Yianilos, "Feature-based face recognition using mixture-distance," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings CVPR '96*, 209 – 216 (1996).

[97] R. Brunelli and T. Poggio, "Face recognition through geometrical features," in *Computer Vision — ECCV'92*, G. Sandini, Ed., *Lecture Notes in Computer Science* **588**, 792–800, Springer Berlin Heidelberg (1992).

[98] J. Križaj, V. Štruc, and N. Pavešić, "Adaptation of sift features for robust face recognition," in *Image Analysis and Recognition*, A. Campilho and M. Kamel,

Eds., *Lecture Notes in Computer Science* **6111**, 394–404, Springer Berlin Heidelberg (2010).

[99] W. Ouarda, H. Trichili, A. Alimi, and B. Solaiman, "Face recognition based on geometric features using support vector machines," in *6th International Conference of Soft Computing and Pattern Recognition SoCPaR*, 89 – 95 (2014).

[100] L. Lang and W. Gu, "Study of face detection algorithm for real-time face detection system," in *Second International Symposium on Electronic Commerce and Security, ISECS '09*, **2**, 129–132 (2009).

[101] L. Zhang and P. Lenders, "Knowledge-based eye detection for human face recognition," in *Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, **1**, 117 – 120 (2000).

[102] O. Jesorsky, K. J. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," in *Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, *AVBPA '01*, 90 – 95, Springer-Verlag, (London, UK, UK) (2001).

[103] S. Kherchaoui and A. Houacine, "Face detection based on a model of the skin color with constraints and template matching," in *International Conference on Machine and Web Intelligence ICMWI*, 469 – 472 (2010).

[104] N. Tiwari, N. K. Mittal, and S. G. Kerhalker, "Automatic face detection in frontal face color images," *International Journal of Scientific Engineering and Technology* **2**(7), 670 – 674 (2013).

[105] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 34 – 58 (2002).

[106] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience* **3**, 71–86 (1991).

[107] C. Liu, "Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance," *IEEE Transactions on Pattern*

Analysis and Machine Intelligence **28**, 725–737 (2006).

[108] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 711 – 720 (1997).

[109] P. Robinson and W. Clarke, "Comparison of principal component analysis and linear discriminant analysis for face recognition," in *AFRICON*, 1 – 6, IEEE (2007).

[110] D. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(8), 831 – 836 (1996).

[111] C. H. Park and H. Park, "Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition," *SIAM Journal on Matrix Analysis and Applications* **27**(1), 87–102 (2005).

[112] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation* **12**, 2385–2404 (2000).

[113] K. Delac, M. Grgic, and S. Grgic, "Image compression effects in face recognition systems," in *Face recognition*, K. Delac and M. Grgic, Eds., InTech (2007).

[114] A. M. Ford, *Relationships between image quality and still image compression.* PhD thesis, PhD thesis, Image Technology Research Group, University of Westminster (1997).

[115] A. Antoniou, *Digital filters: analysis and design*, McGraw-Hill (1979).

[116] X. Ding and C. Fang, "Discussions on some problems in face recognition," in *Advances in Biometric Person Authentication*, S. Li, J. Lai, T. Tan, G. Feng, and Y. Wang, Eds., *Lecture Notes in Computer Science* **3338**, 47 – 56, Springer Berlin Heidelberg (2005).

[117] L. Xu, "Issues in video analytics and surveillance systems: Research / prototyping vs. applications / user requirements," in *IEEE Conference on Advanced Video and Signal Based Surveillance, 2007. AVSS 2007*, 10 – 14 (2007).

[118] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, and V. Placidi, "Customers' activity recognition in intelligent retail environments," in *New Trends in Image Analysis and Processing – ICIAP 2013*, A. Petrosino, L. Maddalena, and P. Pala, Eds., *Lecture Notes in Computer Science* **8158**, 509–516, Springer Berlin Heidelberg (2013).

[119] C.-F. Shu, A. Hampapur, M. Lu, L. Brown, J. Connell, A. Senior, and Y. Tian, "IBM smart surveillance system (S3): an open and extensible framework for event based surveillance," in *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005. AVSS 2005*, 318–323 (2005).

[120] S. Yu, B. Li, Q. Zhang, C. Liu, and M. Q.-H. Meng, "A novel license plate location method based on wavelet transform and {EMD} analysis," *Pattern Recognition* **48**(1), 114 – 125 (2015).

[121] S. Patil and K. Talele, "Suspicious movement detection and tracking based on color histogram," in *International Conference on Communication, Information Computing Technology (ICCICT)*, 1 – 6 (2015).

[122] S. Kim, J. Shi, A. Alfarrarjeh, D. Xu, Y. Tan, and C. Shahabi, "Real-time traffic video analysis using intel viewmont coprocessor," in *Databases in Networked Information Systems*, A. Madaan, S. Kikuchi, and S. Bhalla, Eds., *Lecture Notes in Computer Science* **7813**, 150–160, Springer Berlin Heidelberg (2013).

[123] A. Samama, "Innovative video analytics for maritime surveillance," in *2010 International Waterside Security Conference WSS*, 1 – 8 (2010).

[124] M. Razali, "Detection and classification of moving object for smart vision sensor," in *ICTTA '06, 2nd, Information and Communication Technologies*, **1**, 733 – 737 (2006).

[125] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE* **90**, 1151 – 1163 (2002).

[126] M. Hedayati, W. Zaki, and A. Hussain, "Real-time background subtraction for video surveillance: From research to reality," in *6th International Colloquium on Signal Processing and Its Applications (CSPA)*, 1 – 6 (2010).

[127] A. Tartakovsky and J. Brown, "Adaptive spatial-temporal filtering methods for clutter removal and target tracking," *IEEE Transactions on Aerospace and Electronic Systems* **44**, 1522 – 1537 (2008).

[128] M. Paul, S. M. E. Haque, and S. Chakraborty, "Human detection in surveillance videos and its applications - a review," *EURASIP Journal on Advances in Signal Processing* **2013**(1) (2013).

[129] M. Tsuchiya and H. Fujiyoshi, "Evaluating feature importance for object classification in visual surveillance," in *18th International Conference on Pattern Recognition, 2006. ICPR*, **2**, 978 – 981 (2006).

[130] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, **1**, 511– 518 (2001).

[131] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.* **38** (2006).

[132] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, 1 – 8 (2007).

[133] M.Han, A. Sethi, W.Hua, and Y. Gong, "A detection-based multiple object tracking method," in *International Conference on Image Processing, 2004. ICIP '04*, **5**, 3065–3068 (2004).

[134] R. Kaucic, A. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs, "A unified framework for tracking through occlusions and across sensor gaps," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, **1**, 990–997 (2005).

[135] L. Shengnan, S. Huansheng, C. Hua, and W. Guofeng, "A point-based tracking algorithm for vehicle trajectories in complex environment," in *Fifth International Conference on Intelligent Systems Design and Engineering Applications ISDEA*, 69–73 (2014).

[136] D. H. Hung, H. Saito, and G.-S. Hsu, "Detecting fall incidents of the elderly based on human-ground contact areas," in *2nd IAPR Asian Conference on Pattern Recognition ACPR*, 516–521 (2013).

[137] T. B. Moeslund, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding* **104**, 90 – 127 (2006).

[138] T. A.-Q. Tawiah, *Video content analysis for automated detection and tracking of humans in CCTV surveillance applications.* PhD thesis, School of Engineering and Design, Brunel University (2010).

[139] H. M. Dee and S. A. Velastin, "How close are we to solving the problem of automated visual surveillance?," *Machine Vision and Applications* **19**(5-6), 329–343 (2008).

[140] "PETS 2006 dataset," in *Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance PETS 2006*, J. M. Ferryman, Ed., James M. Ferryman and IEEE.

[141] "CAVIAR: Context aware vision using image-based active recognition," project, EC's Information Society Technology's programme project IST 2001 37540 (2005).

[142] "Video analysis and content extraction VACE," tech. rep., Disruptive Technology Office, Information Exploitation Research Division, US (2007).

[143] "Guidelines for trec video retrieval evaluation TRECVID," tech. rep., NIST (2015).

[144] S. Munder and D. Gavrila, "An experimental study on pedestrian classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(11), 1863 – 1868 (2006).

[145] M.Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes VOC dataset and challenge," (2009).

[146] P. Korshunov and W. T. Ooi, "Video quality for face detection, recognition, and tracking," *ACM Transactions on Multimedia Computing, Communications and Applications* **7**, 14:1–14:21 (2011).

[147] C. Poynton, "Raster images," in *Digital video and HDTV : Algorithms and Interfaces*, ch. 1, 3 – 17, Morgan Kaufmann (2003).

[148] D. Austerberry, "Video formats," in *The technology of video and audio streaming*, ch. 4, 52–77, Elsevier, 2 ed. (2005).

[149] M. Ghanbari, "Video basics," in *Standard codecs: image compression to advanced video coding*, *IEE Telecommunications 49*, ch. 2, 9 – 24, The institution of electrical engineers, London.

[150] "ITU. encoding parameters of digital television for studios," Rec. ITU-R BT.601-4 (1994).

[151] M. Bosch, F. Zhu, and E. Delp, "Segmentation-based video compression using texture and motion models," *IEEE Journal of Selected Topics in Signal Processing* **5**, 1366–1377 (2011).

[152] G. Sullivan and T. Wiegand, "Video compression - from concepts to the H.264/AVC standard," *Proceedings of the IEEE* **93**, 18–31 (2005).

[153] T. Sikora, "Trends and perspectives in image and video coding," *Proceedings of the IEEE* **93**, 6–17 (2005).

[154] "ITU. information technology - lossless and near lossless compression of continuous tone still images Baseline," Rec. ITU-T T.87 (1998).

[155] N. R. Axford, "Digital image processing and manipulation," in *The Manual of Photography, photographic and digital imaging*, G. G. A. R. E. Jacobson, S. F. Ray and N. R. Axford, Eds., ch. 26, 428 – 446, Elsevier Science Ltd, 9 ed. (2000).

[156] M. Nadenau, J. Reichel, and M. Kunt, "Wavelet-based color image compression: exploiting the contrast sensitivity function," *IEEE Transactions on Image processing* **12**(1), 58–70 (2003).

[157] P. Symes, "What is compression?," in *Video compression demystified*, ch. 1, 1 – 14, McGraw-Hill (2001).

[158] M. Ghanbari, "Principles of video compression," in *Standard codecs: image compression to advanced video coding*, *IEE Telecommunications 49*, ch. 3, 25 – 62, Standard codecs: image compression to advanced video coding.

[159] D. Le Gall, "MPEG: A video compression standard for multimedia applications," *Communications of the ACM - Special issue on digital multimedia systems* **34**, 46–58 (1991).

[160] H. Jozawa, K. Kamikura, A. Sagata, H. Kotera, and H. Watanabe, "Two-stage motion compensation using adaptive global mc and local affine mc," *IEEE Transactions on Circuits and Systems for Video Technology* **7**, 75–85 (1997).

[161] J. H. Kim, A. Ortega, P. Yin, P. Pandit, and C. Gomila, "Motion compensation based on implicit block segmentation," in *15th IEEE International Conference on Image Processing, ICIP 2008*, 2452–2455 (2008).

[162] P. Eisert, T. Wiegand, and B. Girod, "Model-aided coding: a new approach to incorporate facial animation into motion-compensated video coding," *IEEE Transactions on Circuits and Systems for Video Technology* **10**(3), 344–358 (2000).

[163] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Perceptual prefiltering for video coding," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, 510–513 (2004).

[164] P. Symes, "Transforms," in *Video compression demystified*, ch. 5, 67 – 90, McGraw-Hill (2001).

[165] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology* **13** (2003).

[166] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with H.264/AVC: tools, performance, and complexity," *IEEE Circuits and Systems Magazine* **4**(1), 7–28 (2004).

[167] P. Symes, "MPEG-2," in *Video compression demystified*, ch. 10, 171 – 192, McGraw-Hill (2001).

[168] E. Allen, S. Triantaphillidou, and R. Jacobson, "Perceptibility and acceptability of JPEG 2000 compressed images of various scene types," *Proc. SPIE 9016* , 90160W–90160W–15 (2014).

[169] P. G. Engeldrum, "Image quality and psychometric scaling," in *Psychometric scaling: A toolkit for imaging systems development*, P. G. Engeldrum, Ed., ch. 1, 1–4, Imcotek Press, Winchester, USA (2000).

[170] P. G. Engeldrum, "The image quality circle," in *Psychometric scaling: A toolkit for imaging systems development*, P. G. Engeldrum, Ed., ch. 2, 5–18, Imcotek Press, Winchester, USA (2000).

[171] S. Triantaphillidou, *Aspects of Image Quality in the Digitisation of Photographic collections*. PhD thesis, Image Technology Research Group, University of Westminster (2001).

[172] E. A. Fedorovskaya, H. de Ridder, and F. J. J. Blommaert, "Chroma variations and perceived quality of color images of natural scenes," *Color Research*

& Application **22**(2), 96–110 (1997).

[173] H. de Ridder, F. J. J. Blommaert, and E. A. Fedorovskaya, "Naturalness and image quality: chroma and hue variation in color images of natural scenes," (1995).

[174] J. Roufs, "Perceptual image quality : concept and measurement," *Philips Journal of Research* **47**(1), 35–62 (1992).

[175] M. A. Arbib and A. R. Hanson, "Vision, brain, and cooperative computation: an overview," in *Vision, Brain, and Cooperative Computation*, M. A. Arbib and A. R. Hanson, Eds., 1–83, MIT Press (1987).

[176] Pcimag.com, "Color: The silent language," (2002).

[177] M. M. Aslam, "Are You Selling the Right Colour? A Crosscultural Review of Colour as a Marketing Cue," *Journal of Marketing Communications* **12**, 15–30 (2006).

[178] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters* **9**, 81–84 (2002).

[179] P. C. Cosman, R. M. Gray, and R. A. Olshen, "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy," *Proceedings of The IEEE* **82**, 919–932 (1994).

[180] M. D. Fairchild and G. M. Johnson, "The iCAM framework for image appearance, image differences, and image quality," *Journal of Electronic Imaging* **13**, 126–138 (2004).

[181] B. Girod, "Psychovisual aspects of image processing: What's wrong with mean squared error?," in *Proceedings of the Seventh Workshop on Multidimensional Signal Processing*, 2–2 (1991).

[182] P. Teo and D. Heeger, "Perceptual image distortion," in *IEEE International Conference Image Processing, Proceedings, ICIP-94*, **2**, 982–986 (1994).

[183] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing* **70**(3), 177 – 200 (1998).

[184] K. J. Kim, B. Kim, R. Mantiuk, T. Richter, H. Lee, H.-S. Kang, J. Seo, and K. H. Lee, "A comparison of three image fidelity metrics of different computational principles for JPEG2000 compressed abdomen ct images.," *IEEE Transactions on Medical Imaging* **29**, 29 (2010).

[185] S. Pasqualini, F. Fioretti, A. Andreoli, and P. Pierleoni, "Comparison of H.264/AVC, H.264 with AIF, and AVS based on different video quality metrics," in *International Conference on Telecommunications*, 190 – 195 (2009).

[186] J. Farrell, "Image quality evaluation," in *MacDonald, L. W., Luo, M. R., Colour image science - Exploring digital media*, 285–314, John Wiley and Sons (1999).

[187] D. Silverstein and J. Farrell, "The relationship between image fidelity and image quality," in *International Conference on Image Processing, 1996. Proceedings*, **1**, 881–884 (1996).

[188] I. A.-A. Abdul-Jabbar, "Image processing for face recognition rate enhancement," *International Journal of Advanced Science and Technology* **64**, 1 (2014).

[189] J. Shermina, "Illumination invariant face recognition using discrete cosine transform and principal component analysis," in *2011 International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT)*, 826–830 (2011).

[190] B. Wang, W. Li, W. Yang, and Q. Liao, "Illumination normalization based on weber's law with application to face recognition," *IEEE Signal Processing Letters* **18**(8), 462–465 (2011).

[191] H. Sellahewa and S. Jassim, "Image-quality-based adaptive face recognition," *IEEE Transactions on Instrumentation and Measurement* **59**, 805–813 (2010).

[192] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR '09*, 597–604 (2009).

[193] D.-H. Liu, K.-M. Lam, and L.-S. Shen, "Illumination invariant face recognition," *Pattern Recogn.* **38**, 1705–1716 (2005).

[194] "ITU. methodology for the subjective assessment of the quality of television pictures," Rec. ITU-R BT.500-11 (2002).

[195] H. Shi, Y. Zou, J. Li, H. Chen, and X. Ma, "Optical design and optimization of panoramic objective lens system," in *International Conference on Optoelectronics and Microelectronics (ICOM)*, 328–330 (2012).

[196] R. E. Jacobson and S. Triantaphillidou, "Metric approaches to image quality," in *Colour image science Exploiting Digital media*, L. W. MacDonald and M. R. Luo, Eds., ch. 18, 371–392, John Wiley and Sons Ltd (2002).

[197] P. G. J. Barten, "Evaluation of subjective image quality with the square-root integral method," *Journal of the Optical Society of America A* **7**, 2024 – 2031 (1990).

[198] R. B. Jenkin, S. Triantaphillidou, and M. A. Richardson, "Effective pictorial information capacity as an image quality metric," in *Proc. SPIE 6494 Image Quality and System Performance IV, 64940O*, (2007).

[199] S. Triantaphillidou, "Introduction to image quality and system performance," in *The Manual of Photography Tenth Edition*, E. Allen and S. Triantaphillidou, Eds., ch. 19, 345–363, Elsevier Ltd (2011).

[200] L. Janowski, M. Leszczuk, M.-C. Larabi, and A. Ukhanova, "Recognition tasks," in *Quality of Experience*, S. Möller and A. Raake, Eds., *T-Labs Series in Telecommunication Services*, 383–394, Springer International Publishing (2014).

[201] G. Kuhn, M. Oliveira, and L. A. Fernandes, "An efficient naturalness-preserving image-recoloring method for dichromats," *IEEE Transactions on Visualization and Computer Graphics* **14**, 1747–1754 (2008).

[202] J. Zhao, "Effect of JPEG2000 compression on fingerprint image quality." MSc degree in Digital and Photographic course at the University of Westminster (2006).

[203] R. W. G. Hunt, "Trichomatic colour reproduction and the additive principle," in *The reproduction of colour*, ch. 2, 9 – 17, John Wiley and Sons Ltd (2004).

[204] H.-J. Liu, M.-J. Liaw, and H.-P. D. Shieh, "Empirical tone reproduction curve equation applied for color characterization of LCDs," in *Proceedings of the 5th Asian Symposium on Information Display, 1999. ASID '99*, 111–114 (1999).

[205] C. E. K. Mees, "L. A. Jones and his work on photographic sensitometry," *Image, Journal of Photography of George Eastman House* **3**(5), 34 (1954).

[206] G. G. Attridge, "Sensitometry," in *The Manual of Photography, photographic and digital imaging, Ninth Edition*, R. E. Jacobson, S. F. Ray, G. G. Attridge, and N. R. Axford, Eds., ch. 15, 218 – 246, Elsevier Science Ltd (2000).

[207] C. Liu and M. D. Fairchild, "Measuring the relationship between perceived image contrast and surround illumination," *Color and Imaging Conference* (1), 282–288 (2004).

[208] M. Fairchild, *Color Appearance Models*, Addison-Wesley (2005).

[209] M. D. Fairchild, "Human colour vision," in *Color appearance models*, *Imaging Science and Technology, Editor M. A. Kriss*, ch. 1, 1–34, Wiley, IS & T Series, 2 ed. (2005).

[210] R. W. G. Hunt, *Measuring colours*, Ellis Horwood, 2 ed. (1991).

[211] "CIE. international lighting vocabulary," CIE publication no.17.4 (1987).

[212] T. Young, "The Bakerian lecture: On the theory of light and colours," *Philosophical Transactions of the Royal Society of London* **92**, 12–48 (1802).

[213] M. S. Longair, "Maxwell and the science of colour," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **366**(1871), 1685–1696 (2008).

[214] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour difference formula: CIEDE2000," *Color Research and Application* **26**(5), 340 – 350 (2001).

[215] M. D. Fairchild and G. M. Johnson, "The iCAM Framework for Image Appearance, Image Differences, and Image Quality," *Journal of Electronic Imaging* **13**, 126–138 (2004).

[216] J. Kuang, G. M. J., and M. D. Fairchild, "iCAM06: A refined image appearance model for HDR image rendering," *Journal of Visual Communication and Image Representation* **18**(5), 406 – 414 (2007). Special issue on High Dynamic Range Imaging.

[217] A. Calabria and M. Fairchild, "Perceived image contrast and observer preference: I. the effects of lightness, chroma, and sharpness manipulations on contrast perception," *The Journal of Imaging Science and Technology* **47**(6), 479–493 (2003).

[218] I. A. Cunningham and A. Fenster, "A method for modulation transfer function determination from edge profiles with correction for finite-element differentiation.," *Medical Physics* **14**, 533–537 (1987).

[219] N. R. Axford, "Theory of image formation," in *The Manual of Photography, photographic and digital imaging*, R. E. Jacobson, S. F. Ray, G. G. Attridge, and N. R. Axford, Eds., ch. 24, 393–412, Elsevier Science Ltd, 9 ed. (2000).

[220] K. J. Barnard, G. D. Boreman, A. E. Plogstedt, and B. K. Anderson, "Modulation-transfer function measurement of sprite detectors: sine-wave response.," *Applied Optics* **31**, 144–147 (1992).

[221] G. D. Boreman, "Transfer function techniques," in *Handbook of optics: Devices, measurements, and properties*, M. Bass, Ed., ch. 32, McGraw-Hill

(1995).

[222] B. Theron, M. El-Desouki, F. Aljekhedab, M. Alayed, and M. Alsawadi, "Choice of spatial resolution measurement methods to implement," in *Electronics, Communications and Photonics Conference (SIECPC), 2013 Saudi International*, 1–5 (2013).

[223] K. Masaoka, T. Yamashita, Y. Nishida, and M. Sugawara, "Modified slanted-edge method and multidirectional modulation transfer function estimation.," *Optics Express* **22**, 6040–6046 (2014).

[224] S. A. Klein, "Measuring, estimating, and understanding the psychometric function: a commentary.," *Perception and Psychophysics* **63**, 1421–1455 (2001).

[225] P. F. Judy, "The line spread function and modulation transfer function of a computed tomographic scanner," *Medical Physics* **3**(4), 233–236 (1976).

[226] E. Heynacher and F. Kober, "Resolving power and contrast," Tech. Rep. 51, Carl Zeiss, Germany (1976).

[227] "Photography –Electronic still-picture cameras–Resolution measurements," tech. rep., ISO 12233:2000 (2000).

[228] R. Jenkin, "Noise, sharpness, resolution and information," in *The Manual of Photography*, E. Allen and S. Triantaphillidou, Eds., ch. 24, 433 – 456, Elsevier Ltd (2011).

[229] "Photography –Root mean square granularity of photographic films–Method measurement," ISO, ISO 10505:2009 (2009).

[230] M. D. Fairchild and R. S. Berns, "Image color appearance specification through extension of CIELAB," *Color Research & Application Application* **18**(3), 178–190 (1993).

[231] G. M. Johnson and M. D. Fairchild, "A top down description of S - CIELAB and CIEDE2000," *Color Research and Application* **28**(6), 425–435 (2003).

[232] "CIE standard colorimetric observers," standard, BS ISO CIE 10527 (1991).

[233] M. D. Fairchild, "Colorimetry," in *Color appearance models*, *Imaging Science and Technology, Editor M. A. Kriss*, ch. 3, 53–82, Wiley, IS & T Series, 2 ed. (2005).

[234] "Applied image QA-62 Test Chart," chart, Applied Imaging Inc (2014).

[235] R. M. Boynton and C. L. Bartleson, "Optical radiation measurements," in *Visual measurements part II*, **5**, ch. 7 - 9, Academic press, New York, USA (1984).

[236] D. Williams and P. D. Burns, "Measuring and managing digital image sharpening," *Proc. IS&T 2008, Archiving Conference* , 89–93 (2008).

[237] G. Gescheider, *Psychophysics: The fundamendals*, ix. Lawrence Erlbaum Associates, 3 ed. (1997).

[238] P. G. Engeldrum, "The process of scaling and some practical hints," in *Psychometric scaling: A toolkit for imaging systems development*, P. G. Engeldrum, Ed., ch. 3, 19–42, Imcotek Press, Winchester, USA (2000).

[239] E. H. Adelson, "Lightness perception and lightness illusions," in *The New Cognitive Neurosciences*, 339–351, MA: MIT Press, 2 ed. (2000).

[240] S. S. Stevens, "On the theory of scales of measurement," *Science* **103**(2684), 677–680 (1946).

[241] P. G. Engeldrum, "Measurement scales," in *Psychometric scaling: A toolkit for imaging systems development*, P. G. Engeldrum, Ed., ch. 4, 43–52, Imcotek Press, Winchester, USA (2000).

[242] W. H. Ehrenstein and A. Ehrenstein, "Psychophysical methods," in *Modern Techniques in Neuroscience Research*, U. Windhorst and H. Johansson, Eds., 1211–1241, Springer Berlin Heidelberg (1999).

[243] P. G. Engeldrum, "Thresholds and just-noticeable differences," in *Psychometric scaling: A toolkit for imaging systems development*, P. G. Engeldrum,

Ed., ch. 5, 53–78, Imcotek Press, Winchester, USA (2000).

[244] G. Ekman and U. Gustafsson, "Threshold values and the psychophysical function in brightness vision," *Vision Research* **8**(6), 747 – 758 (1968).

[245] H. Resnikoff, "On the psychophysical function," *Journal of Mathematical Biology* **2**(3), 265–276 (1975).

[246] B. C. Duchaine and A. Weidenfeld, "An evaluation of two commonly used tests of unfamiliar face recognition.," *Neuropsychologia* **41**(6), 713–720 (2003).

[247] C. Wilkinson and R. Evans, "Are facial image analysis experts any better than the general public at identifying individuals from CCTV images?," *Science and Justice Justice* **49**, 191–196 (2009).

[248] R. Dugad and N. Ahuja, "A fast scheme for image size change in the compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology* **11**(4), 461–474 (2001).

[249] "Subjective video quality assessment methods for multimedia applications," Rec. ITU-R P.910, ITU (1999).

[250] K. Bashir, T. X., and G. Shaogang, "Feature selection on gait energy image for human identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2008*, 985 – 988 (2008).

[251] D. Fell, "Private communications - TfL." (2011).

[252] TFL, "Transport for london - CCTV. Available from: https://www.tfl.gov.uk/corporate/privacy-and-cookies/cctv," (2015).

[253] R. Kemp, G. Pike, P. White, and A. Musselman, "Perception and recognition of normal and negative faces: the role of shape from shading and pigmentation cues," *Perception* **25**, 37–52 (199).

[254] E. W. Jin, B. W. Keelan, J. Chen, J. B. Phillips, and Y. Chen, "Softcopy quality ruler method: implementation and validation," in *Proc. SPIE 7242,*

Image Quality and System Performance VI, 724206, San Francisco, S. P. Farnand and F. Gaykema, Eds. (2009).

[255] "CASTBUS 2012," Dataset, Home Office Centre for Applied Sience and Technology, UK (2012).

[256] C. Poynton, "DV Compression," in *Digital video and HDTV : Algorithms and Interfaces*, 461 – 471, M. Kaufmann (2003).

[257] C. Poynton, "MPEG2 video compression," in *Digital video and HDTV : Algorithms and Interfaces*, 474 – 496, M. Kaufmann (2003).

[258] A. Tsifouti, M. M. Nasralla, M. Razaak, J. Cope, J. M. Orwell, M. G. Martini, and K. Sage, "A methodology to evaluate the effect of video compression on the performance of analytics systems," (2012).

[259] S. Ishihara, *Test for colour-blindness*, Kanehara Shuppan Co. Ltd (1972).

[260] F. A. A. Kingdom and N. Prints, "Classifying psychophysical experiments," in *Psychophysics : A practical introduction*, ch. 2, 9 – 37, Elsevier Ltd (2010).

[261] F. A. A. Kingdom and N. Prints, *Psychophysics : A practical introduction*, Elsevier Ltd (2010).

[262] F. A. Wichmann and N. J. Hill, "The psychometric function: I. fitting, sampling, and goodness of fit.," *Perception and Psychophysics* **63**, 1293–1313 (2001).

[263] F. A. A. Kingdom and N. Prints, "Psychometric functions," in *Psychophysics : A practical introduction*, ch. 4, 59 – 118, Elsevier Ltd (2010).

[264] D. Collett, *Modelling binary data*, Chapman & Hall/CRC Texts in Statistical Science, Chapman & Hall (1991).

[265] A. J. Dobson, "Binary variables and logistic regression," in *An introduction to generalised linear models*, C. Chatfield and J. Zidek, Eds., *Texts in Statistical Science*, 120 – 139, Chapman and Hall/CRC (2002).

[266] F. A. Wichmann and N. J. Hill, "The psychometric function: Ii. bootstrap-based confidence intervals and sampling.," *Perception and Psychophysics* **63**, 1314–1329 (2001).

[267] A. Janssen and T. Pauls, "How do bootstrap and permutation tests work?," *The Annals of Statistics* **31**(3), 768 – 806 (2003).

[268] B. W. Keelan, "Just noticeable differences," in *Handbook of Image Quality:Characterization and Prediction*, B. W. Keelan, Ed., ch. 3, 35–46, Marcel Dekker Inc, New York (2002).

[269] A.-N. Spiess and N. Neumeyer, "An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach," *BMC Pharmacology* **10**(1) (2010).

[270] P. J. Hancock, V. V. Bruce, and A. M. Burton, "Recognition of unfamiliar faces.," *Trends in Cognitive Sciences* **4**, 330–337 (2000).

[271] V. V. Bruce, A. M. Burton, and N. Dench, "What's distinctive about a distinctive face?," *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology* **47**, 119–141 (1994).

[272] J. Penry, *Looking at faces and remembering them: a guide to facial identification*, Elek (1971).

[273] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin* **1**, 80–83 (1945).

[274] S. N. Yendrikhovskij, "Image quality: Between science and fiction," in *PICS*, 173–178, The Society for Imaging Science and Technology (1999).

[275] Y. Hu and Z. Wang, "A similarity measure based on hausdorff distance for human face recognition," in *18th International Conference on Pattern Recognition, 2006. ICPR 2006*, **3**, 1131–1134 (2006).

[276] V. A. Mane, R. R. Manza, and K. V. Kale, "The role of similarity measures in face recognition," *International journal of computer science and application*

(2010).

[277] S. Trivedi, "Face recognition using eigenfaces and distance classifiers." A Tutorial (2009).

[278] V. Štruc and N. Pavešić, "Gabor-based kernel partial-least-squares discrimination features for face recognition," *Informatica* **20**, 115 – 138 (2009).

[279] V. Štruc and N. Pavešić, "The complete gabor-fisher classifier for robust face recognition," *EURASIP Journal on Advances in Signal Processing* , 1 – 13 (2010).

[280] B. Sheng, W. Gao, and D. Wu, "An implemented architecture of deblocking filter for H.264/AVC," in *International Conference on Image Processing, 2004. ICIP '04. 2004*, 665–668 (2004).

[281] A. M. Brown, "A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft Excel spreadsheet," *Computer Methods and Programs in Biomedicine* **65**(3), 191 – 200 (2001).

[282] "R core team, R : a language and environment for statistical computing." Foundation for statistical computing, Vienna, Austria (2013).

[283] R. I. Kabacoff, "Regression," in *R in action: Data analysis and graphics with R*, S. Stirling and L. Welch, Eds., ch. 8, 173 – 218, Manning (2011).

[284] J. Fox and S. Weisberg, "Nonlinear regression and nonlinear least squares in R, An Appendix to an R companion to applied regression, 2nd edition," in *An R Companion to Applied Regression*, SAGE Publications, Inc, 2 ed. (2011).

[285] H. G. Dietz and P. S. Eberhart, "ISO-less?," *Proc. SPIE 9404* , 94040L–94040L–14 (2015).

[286] K. Spaulding, R. Joshi, and G. Woolfe, "Using a residual image formed from a clipped limited color gamut digital image to represent an extended color gamut digital image," (2001). US Patent 6,301,393.

[287] R. Joshi, K. Spaulding, and G. Woolfe, "Method and apparatus to represent an extended color gamut digital image using a residual image," (2001). EP Patent App. EP20,000,109,418.

[288] "VideoLAN VLC, GNU-General Public License Version 2."

[289] "nVLC, Code Project website, Roman Ginzburg, GNU-General Public License (GPLv3)."

[290] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters* **31**(8), 651 – 666 (2010). Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)19th International Conference in Pattern Recognition (ICPR).

[291] M. J. Crawley, "Proportion data," in *Statistics: An introduction using R*, ch. 14, 248–262, John Wiley & Sons Ltd (2005).

[292] R. I. Kabacoff, "Generalized linear models," in *R in action: Data analysis and graphics with R*, S. Stirling and L. Welch, Eds., ch. 13, 313–330, Manning (2011).

[293] G. Rodriguez, "Lecture notes on generalized linear models," tech. rep., Princeton University (2015). Lecture notes on generalized linear models.

[294] M. Grgic, K. Delac, and S. Grgic, "SCface – surveillance cameras face database," *Multimedia Tools and Applications* **51**(3), 863–879 (2011).

[295] "Face recognition home page. http://www.face-rec.org/databases/," (2015).

[296] R. Gross, "Face databases," in *Handbook of Face Recognition*, 301–327, Springer New York (2005).

[297] A. Maalouf, M. C. Larabi, and D. Nicholson, "Offline quality monitoring for legal evidence images in video surveillance application," *Journal of Multimedia Tools and Applications* **73**, 189 – 218 (2014).

[298] J. Y. Park, *Evaluation of changes in image appearance with changes in displayed image size*. PhD thesis, University of Westminster (2014).

[299] "BSI - BS EN 61966-4. Multimedia Systems and Equipment - Colour Measurement and Management - Part 4: Equipment Using Liquid Display Panels," (2000).

[300] J. D. Onken, B. S. Caldwell, and S. A. Murray, "Human factors in displays," in *Measurement, instrumentation and sensors handbook. Mechanical, Thermal, and Radiation measurement*, J. G. Webster and H. Eren, Eds., ch. 96, 1–17, CRC Press Taylor & Francis Group, 2 ed. (2014).

[301] S. Triantaphillidou, "Tone reproduction," in *The Manual of Photography*, E. Allen and S. Triantaphillidou, Eds., ch. 21, 377–392, Elsevier Ltd (2011).

[302] W. Mokrzycki and M. Tatol, "Color difference E - a survey," *Machine Graphics and Vision* **20**(4) (2011).

[303] G. Hong and M. R. Luo, "Perceptually-based color difference for complex images," (2002).

[304] S. Farnand, "The effect of image content on color difference perceptibility," in *Fourth Color Imaging Conference: Color Science, Systems, and Applications*, 101–104 (1996).

[305] L. Yang, M. Egawa, M. Akimoto, and M. Miyakawa, "An imaging colorimeter for noncontact skin color measurement," *Optical Review* **10**(6), 554–561 (2003).